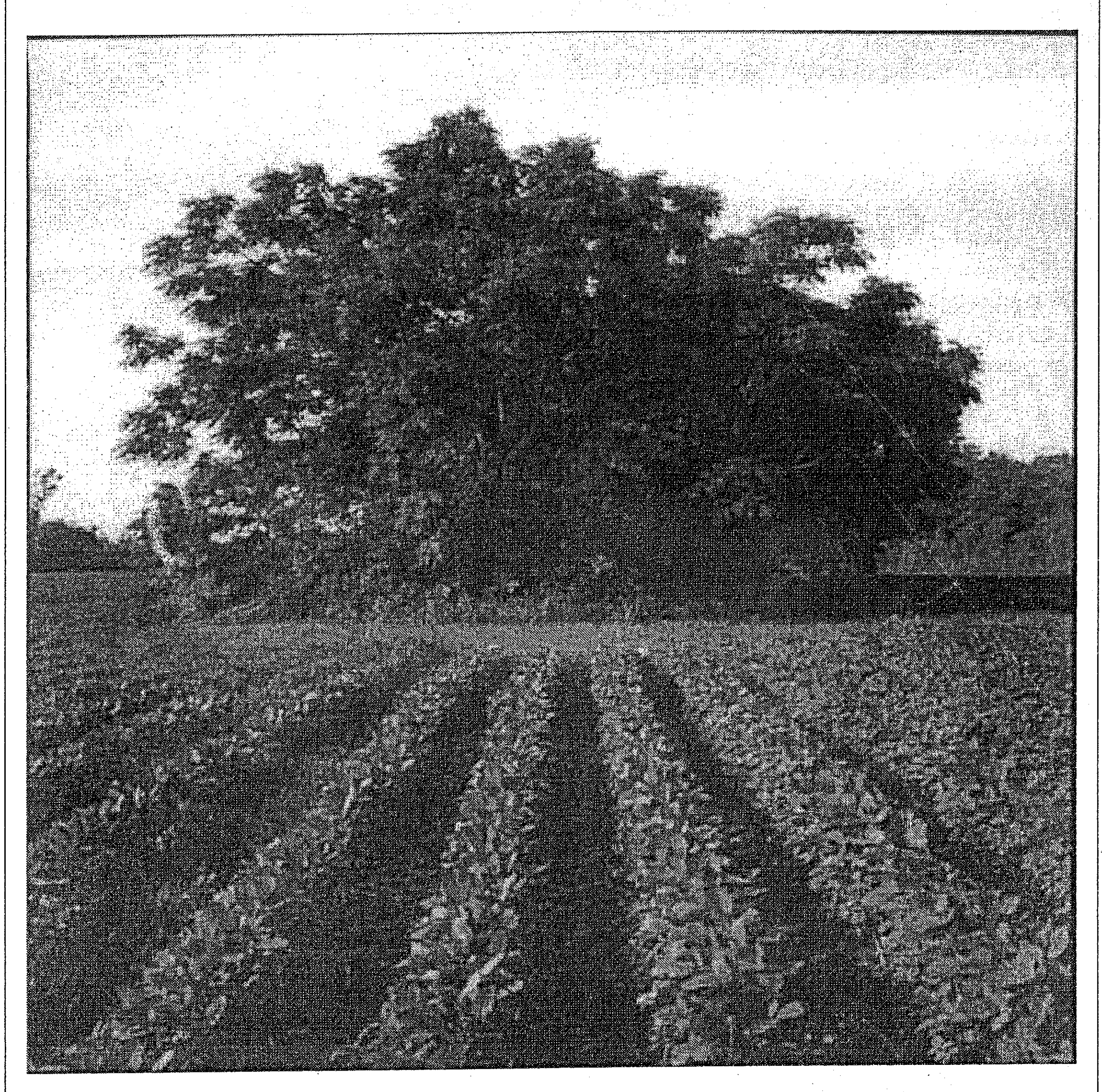
ESTADÍSTICA GENERAL



Manual de clases
Teóricas y Prácticas

Para las carreras de Agronomía,
Ciencias Ambientales y Gestión de Agroalimentos

CÁTEDRA DE MÉTODOS CUANTITATIVOS APLICADOS
FACULTAD DE AGRONOMÍA
Universidad de Buenos Aires

Centro de Impresiones, **Edición 2007**

Cátedra de Métodos Cuantitativos Aplicados - FAUBA

Ubicación
Pabellón Wernicke (Ver plano en http://www.agro.uba.ar/fauba/ubicacion/plano.htm)
Teléfono: 4524-8077.
Secretaria: Ana Inés Garaño (e-mail: garano@agro.uba.ar)

Docentes

Ing. Agr. Susana B. Perelman (Profesora Asociada- Coordinadora de la cátedra)
Ing. Agr. Norberto J. Bartoloni (Profesor Asociado)
Lic. Mirta González (Profesora Asociada)
Ing. Agr. William Batista (Profesor Adjunto)
Ing. Agr. María V. López (Profesora Adjunta)
Lic. Olga S. Filippini (Profesora Adjunta)
Lic. Víctor Brescia (Profesor Adjunto)
Lic. María C. Fabrizio (Profesora Adjunta)
Ing. Agr. Rosa T. Boca (Jefe de Trabajos Prácticos)
Ing. Civ. Diana Giorgini (Jefe de Trabajos Prácticos)
Ing. Agr. Juan P. Guerschman (Jefe de Trabajos Prácticos)
Ing. Agr. Gustavo A. Sznaider (Jefe de Trabajos Prácticos)
Dr. Karina Hodara (Jefe de Trabajos Prácticos)
Lic. Roxana Aragón (Jefe de Trabajos Prácticos)
Ing. Agr. Laura Puhl (Jefe de Trabajos Prácticos)
Ing. Agr. Pablo A. Cipriotti (Jefe de Trabajos Prácticos)
Ing. Civ. Victor Vergez (Ayudante de Primera)
Ing. Agr. Lucas Garibaldi (Ayudante de Primera)
Ing. Agr. Gonzalo Grigera (Ayudante de Primera)
Lic. Fernando Biganzoli (Ayudante de Primera)
Ing. Agr. Pedro Tognetti (Ayudante de Primera)
Lic. Marcos Texeira (Ayudante de Primera)
Lic. Gustavo Vázquez (Ayudante de Primera)
Lic. Martín Cristian Poveda (Ayudante de Primera)
Lic. Alexis Cerezo (Ayudante de Primera)
Sra. Eva Florio (Ayudante de Segunda)
Sr. Sebastián Giedzinski (Ayudante de Segunda)
Sra. Erika Hirschowitz Graus (Ayudante de Segunda)
Sr. Facundo Gallo (Asistente alumno ad Honorem)

Cursos dictados por la Cátedra

Carreras de Agronomía, Ciencias Ambientales y Gestión de Agroalimentos

Estadística general (materia cuatrimestral, 5 hs. semanales, en ambos cuatrimestres)
Modelos Estadísticos (materia bimestral, 6hs. semanales, en 1er. y 3er. bimestres)
Introducción a la Programación para las Ciencias Ambientales (materia bimestral, 5 horas semanales, 1er bimestre)

Licenciatura en Administración

Estadística I (materia cuatrimestral, 6 hs. semanales, 2° cuatrimestre)
Econometría (materia cuatrimestral, 4 hs. semanales, 1er. cuatrimestre)
Estadística para Administradores (bimestral, 6 hs. semanales, 1er. bimestre)

Carrera técnica en Floricultura

Estadística (materia bimestral, 2 hs. semanales, 2° bimestre)

Cursos de intensificación y especialización

Estadística aplicada a la protección vegetal
Colección e Interpretación de datos

Escuela para Graduados Alberto Soriano (programas de Maestría y Doctorado)

Estadística Aplicada (Maestría en Agronegocios)
Estadística Aplicada a la Investigación Biológica (Maestría en Biometría y Mejoramiento)
Análisis Multivariado para las Ciencias Biológicas (Maestría en Biometría y Mejoramiento)
Genética de Poblaciones y Evolución (Maestrías de Producción Vegetal y Recursos Naturales)
Introducción a la Estadística (Maestría en Economía Agraria)
Teoría Estadística (Maestría en Biometría y Mejoramiento)
Análisis de la Heterogeneidad de la Vegetación (Maestría en Recursos Naturales)
Métodos sistémicos para la solución de problemas agronómicos (Maestría en Recursos Naturales)
Análisis de Datos Categóricos (Maestría en Biometría y Mejoramiento)

Foto de tapa: María Zorzon

ESTADÍSTICA GENERAL

Orientación General del Curso

Para intervenir eficazmente en los sistemas que manejan, los profesionales de las ciencias agropecuarias y ambientales deben poseer y desarrollar habilidades relacionadas con la obtención, análisis e interpretación de datos mediante métodos cuantitativos. Estas habilidades son esenciales para interpretar críticamente la información científica y técnica disponible así como para evaluar la estructura y el funcionamiento de los sistemas a manejar y los resultados de las intervenciones realizadas. Una característica intrínseca de los sistemas naturales más o menos intervenidos por el hombre es que su variabilidad espacial y temporal introduce incertidumbre en relación con sus características y su comportamiento. En este sentido, los objetivos de este curso son desarrollar en los alumnos:

- la conciencia de esa incertidumbre y de la necesidad de medirla,
- la capacidad para medirla,
- la habilidad para manejarla para la toma de decisiones y
- la capacidad para leer información técnica publicada en medios especializados con capacidad crítica en relación los aspectos metodológicos.

Duración del curso: 16 semanas
Carga horaria: 5 horas por semana (2 de clases teóricas y 3 de clases prácticas)

Régimen de aprobación

Asistencia mínima obligatoria: 75 %
Evaluaciones:

- Trabajo domiciliario (**10** puntos)
- Examen parcial (**30** puntos)
- Examen integrador (**45** puntos)
- Ejercicios para resolver en clase (parcialitos) (**15** puntos)

Condiciones para aprobar por promoción:
Cumplir con la asistencia mínima obligatoria
Presentar el trabajo de seminario en tiempo y forma
Acumular al menos **70** puntos entre los dos exámenes y los parcialitos y trabajos domiciliarios

Condiciones para obtener la condición de alumno regular:
Cumplir con la asistencia mínima obligatoria
Presentar el trabajo de seminario en tiempo y forma
Acumular al menos **40** puntos entre los dos exámenes y los parcialitos y trabajos domiciliarios

Los alumnos que no aprueben por promoción ni alcancen la regularidad quedarán en condición de libres.

Bibliografía de consulta

- Devore J.L. (2003) *Probabilidad y estadística para ingeniería y ciencias*. 5ta. edición. International Thomson Editores, S. A.
- Wackerly D, W. Mendenhall y R. Scheaffer (2002) *Estadística Matemática con Aplicaciones.* 6ta. ed. Thomson: México
- Pagano, M y K. Gauvreau (2001) *Fundamentos de Bioestadística*. Thomson Learning.
- Steel R y J. Torrie (1988) *Bioestadística: Principios y Procedimientos*. McGraw-Hill/Interamericana de México.
- Spiegel, M.R. y L. Stephens (2002) *Estadística* McGraw-Hill.
- Mendenhall, W. (1990) *Estadística para administradores*. Grupo Editorial Iberoamericana.
- Ya Lun Chou. (1978) *Análisis Estadístico*. Interamericana: México.

INDICE

INTRODUCCIÓN A LA ESTADÍSTICA

Estadística: la ciencia de la obtención y análisis de datos.

En las ciencias agropecuarias y ambientales nos encontramos con situaciones que se presentan como un problema a resolver, un profesional que debe abordar la tarea y un conjunto de herramientas de las cuales podrá valerse para realizar su tarea. Entre estas herramientas se encuentra la Estadística con todo su bagaje teórico y metodológico.

La teoría estadística se apoya en la Matemática de la cual puede considerarse una rama y los métodos estadísticos son las herramientas que el ingeniero puede usar para responder preguntas tales como *¿a cuántas personas debería encuestarse antes de una elección como para poder hacer una predicción válida del resultado de la votación?* o *¿cuál de varios herbicidas es el más recomendable con vistas al control de una determinada maleza de los cultivos de maíz?*

Ahora, ¿cuándo será necesario recurrir a los métodos estadísticos? Los métodos estadísticos serán útiles en todas aquellas situaciones en las cuales deban tomarse decisiones o hacer elecciones o emitir opiniones **bajo incertidumbre**. Es decir, dada una determinada cantidad y calidad de información, debe decidirse el camino a seguir y para ello, la Estadística provee los elementos necesarios para que esas decisiones puedan ser tomadas en forma **racional**. A su vez, el grado de racionalidad de las decisiones estará determinado por la calidad y cantidad de **teoría** y de la calidad y cantidad de métodos de **extracción y análisis de la información** de los que se disponga. En este curso elemental e introductorio, expondremos los elementos básicos de la teoría estadística y de los métodos que se sustentan sobre ella buscando, en todo momento, enfocarlos sobre las aplicaciones prácticas más comunes en la ingeniería agronómica.

La situación más común en la que el ingeniero suele encontrarse es aquella en cual dispone de un conjunto de **datos** extraídos de una masa de información mucho más grande y, probablemente, desconocida y de los cuales debe obtener algún tipo de información específica que responda a sus intereses o interrogantes. Las dudas o interrogantes estarán referidos, la inmensa mayoría de las veces, a una **población** grande – y por "grande" estamos entendiendo que la población está compuesta por una cantidad de unidades inabarcable por parte del analista – y de la cual deberá extraerse una pequeña cantidad de unidades denominada **muestra**.

Podemos definir a la **población** como *un conjunto de elementos físicos o conceptuales acerca de los cuales se desea extraer información a través de uno o más procedimientos*. Por ejemplo, todas las plantas de álamo en explotación comercial en el delta del Paraná.

Por otra parte, una **muestra** es *el conjunto de unidades experimentales realmente observadas o consideradas en un procedimiento de extracción de información*. Ejemplo: un conjunto de 50 plantas de álamo que fueron observadas en una dada localidad del delta del Paraná en un momento determinado.

Finalmente, una **unidad experimental** es *la mínima cantidad de elementos de una población pasibles de ser observados o considerados en un procedimiento de extracción de información*. Ejemplo: cada planta de álamo en explotación comercial en el delta del Paraná.

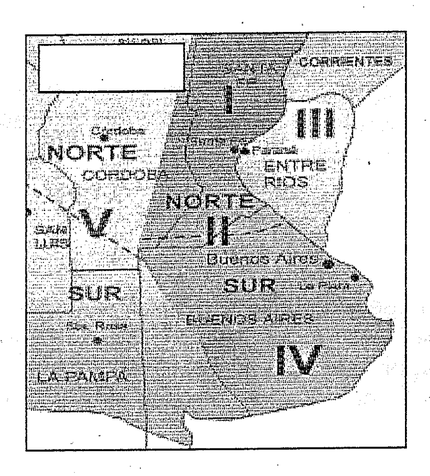
En la población está contenida la masa total de información que sería deseable (pero, quizás, imposible) conocer totalmente. En la muestra, está contenida la porción de información que resulta posible conocer enteramente (los datos) y que servirá para, métodos estadísticos mediante, deducir o conjeturar cómo es todo el resto de la información de la población. A veces, se conoce toda la información contenida en la población. Se trata de poblaciones pequeñas o de poblaciones que, aún siendo grandes, admiten, por una u otra razón ser accedidas por el investigador y, por tanto, en esos casos, puede conocerse toda la información y no es necesario tomar ninguna muestra, sino que, directamente, se realiza un **censo** de toda la población.

Tanto la información contenida en una muestra como la contenida en la población total estarán referidas a una o varias magnitudes o **variables** y pueden ser reducidas o resumidas por una o unas pocas **medidas** que las representen. Es decir, comúnmente, no es necesario conocer todos y cada uno de los valores de las variables de interés sino que bastará con conocer solamente alguna medida resumen de ellos. Las medidas resumen que se calculan a partir de los datos de la muestra se denominan **estadísticas** o **estadísticos** y las correspondientes medidas de dichas variables en la población total, se denominan **parámetros**.

La primera sección del curso se dedicará a exponer los métodos de *organización, presentación* y *descripción* de los datos. Es lo que se denomina **Estadística Descriptiva**. Luego, para el caso más general en que no se puede acceder a toda la información contenida en la población, para tener la posibilidad hacer conjeturas o pronósticos acerca del resto de la información, es decir, acerca de los parámetros, será necesario sentar las bases teóricas de los métodos estadísticos que permiten hacerlo. Por esto, la segunda sección del curso se destinará al estudio de la *teoría de probabilidades*, de las *variables aleatorias*, de los *modelos de probabilidad* más comunes y del *muestreo de distribuciones*. Finalmente, la tercera sección, estará abocada al empleo de los métodos estadísticos que nos permiten hacer conjeturas racionales acerca de los parámetros de la población y, entre ellos, veremos cómo es posible *estimarlos*, y decidir si, a partir de lo que se lee en la muestra, un parámetro es reconocido como perteneciente (o se asume que pertenece) a un determinado conjunto de números, o no. Es lo que se denomina **Estadística Inferencial**.

Como se dijo antes, en general, lo que más interesa conocer no son esos valores concretos de las observaciones muestrales sino los valores de la población total (de todas las observaciones posibles) de la cual provinieron; por ejemplo, la probabilidad de obtener un 5 al arrojar un dado balanceado, o el porcentaje de nacimientos de niñas en la República Argentina durante el próximo año. Una de las preguntas que la **Estadística Inferencial** permite responder es si un conjunto dado de observaciones podrían considerarse como debidas al azar o si, por el contrario, reflejan el efecto de algún factor. Este modo de proceder se ha convertido en el método característico de la ciencia moderna. El científico que descubre fenómenos nuevos, relaciones de dependencia, tendencias o efectos de otro tipo, establece con ellos una **hipótesis de trabajo** y para constatar su validez deberá garantizar de algún modo que los resultados observados no se deben únicamente al azar. Todo estudio de este tipo se basa en la consideración de *muestras aleatorias*, es decir, muestras tales que todas las unidades de la población tengan la misma probabilidad de ser elegidas. Si la población total constara de diversas subpoblaciones parciales bien diferenciadas entre ellas, se tomarán *muestras estratificadas*. Así, por ejemplo, para examinar la calidad panadera de los trigos producidos en la región triguera argentina, que comprende 5 subregiones agroecológicas con distintos escenarios productivos (ver figura 1.1), no podría considerarse como representativa una bolsa de cereal cosechado en la Subregión IV, ni otra proveniente de la Subregión III o de la V; en todo caso, podría ser útil una bolsa que incluyera cereal cosechado en las cinco subregiones. Todavía mejor sería extraer muestras de cada una de las subregiones por separado (Figura 1.1). En los sorteos de lotería se emplean métodos mecánicos para obtener muestras aleatorias. En general, para obtener una muestra aleatoria se enumeran las unidades de la población y a continuación se recurre a una tabla de números aleatorios o a un programa de computadora generador de números aleatorios. Una vez asignado un número a cada unidad perteneciente a la población, se elegirán aquellas cuyos números coincidan con los números obtenidos en el proceso generador aleatorio.

Figura 1.1. Subregiones de producción triguera



La razón fundamental por la cual se debe garantizar un proceso aleatorio de extracción de las muestras reside en el hecho de que podrían subyacer procesos dentro de la población que afecten sistemáticamente a algunas unidades y a otras no, esto es, procesos que afecten a ciertas unidades específicas poseedoras de alguna característica. Si el muestreo es verdaderamente al azar, las chances de ser elegidas serán iguales para todas las unidades, tanto las afectadas por el proceso sistemático como las no afectadas. En cambio, si el muestreo se realiza siguiendo alguna preferencia o idea personal por parte del investigador, podría darse el caso de que su idea o preferencia coincida con el patrón de variación de aquel proceso sistemático y, entonces, podrían resultar elegidas preferentemente las unidades de la población que posean tal característica y los valores numéricos calculados a partir de dicha muestra no reflejarán fielmente lo que pasa en el conjunto total de unidades de la población.

Variables Estadísticas: tipos y escalas de registro.

Las variables en estudio pueden ser de dos tipos: **Cualitativas** o **Cuantitativas**.

Las variables **cualitativas** o atributos *clasifican* o *describen* a las unidades experimentales. Los valores que pueden asumir no constituyen un espacio métrico y, por ello, las operaciones de cálculo no son significativas en ellas. Ejemplos: género, nacionalidad, especie, marca registrada, color, olor, etc.

Las variables **cuantitativas** o numéricas *cuantifican* a las unidades experimentales. Los valores que pueden asumir constituyen un espacio métrico y, por lo tanto, las operaciones de cálculo son significativas en ellas. Ejemplos: cantidad de hojas, número de hijos, kilómetros recorridos, tiempo de vuelo, ingreso familiar, longitud de una espiga, etc. Estas variables cuantitativas pueden a su vez ser distinguidas en **discretas** o **continuas**. Las variables cuantitativas discretas solo pueden asumir una cantidad finita de valores de manera que, entre dos valores cualesquiera, siempre hay huecos. La operación que caracteriza a las variables cuantitativas discretas es la operación de *contar*. Ejemplos: cantidad de materias aprobadas, cantidad de hijos, número de frutos sanos, número de animales marcados, etc. Las variables cuantitativas continuas pueden asumir cualquier valor dentro de un rango dado. La operación que caracteriza a las variables cuantitativas continuas es la operación de *medir*. Se pueden medir longitudes, tiempos, superficies, densidades, volúmenes, sumas de dinero, etc. Ejemplos: peso de un animal al nacer, altura de un árbol, litros de aceite producidos, tiempo de viaje entre dos ciudades, etc.

Para obtener información sobre las variables estadísticas se utilizan diferentes escalas de registro acorde con el tipo de variable. Entre estas escalas de registro se cuentan las escalas nominal, de intervalo y continua.

Escala nominal.

En la escala nominal, las unidades experimentales sólo pueden ser clasificadas en categorías sin ningún ordenamiento ni jerarquía entre ellas. Es aplicable a variables *cualitativas*. Ejemplos: ciudad natal, apellido, color de cabello, color de flor, etc.

Escala de intervalo.

En las escala de intervalo, las unidades experimentales pueden ser clasificadas en categorías las cuales pueden ser ordenadas o jerarquizadas y, además, se pueden establecer diferencias entre categorías. Esta escala es aplicable a las variables *cuantitativas discretas*. Ejemplos: número de personas con empleo, número de plantas con flor, etc.

Escala continua.

En la escala continua, las unidades experimentales pueden ser clasificadas en categorías que pueden ser ordenadas o jerarquizadas y, además, se pueden establecer diferencias entre categorías y las variables pueden tomar cualquier valor real. Sólo es aplicable a las variables *cuantitativas continuas*. Ejemplos: gramos de harina, litros de aceite, tiempo de decantación, etc.

Ejercicios

1.1 Un fabricante de medicamentos veterinarios está interesado en la proporción de animales que padecen infecciones locales cuya condición puede ser controlada por un nuevo producto antibiótico. Se condujo un estudio en el que se tomaron al azar 500 animales que padecían infecciones locales de una estancia de la Pampa Deprimida y se los trató con el medicamento en cuestión. Se encontró que el medicamento controló la enfermedad en el 80% de los animales.

a. *¿Cuál es la* ***población*** *sobre la cual fue conducido este estudio?*
b. *¿Cuál es la* ***muestra*** *que se tomó?*
c. *Identificar el* ***parámetro*** *de interés.*
d. *Identificar la* ***estadística*** *que se estimó y proporcionar su valor.*
e. *¿Se conoce el verdadero valor del parámetro poblacional?*
f. *Si tomamos otra muestra semejante, ¿el valor estimado del nuevo estadístico será idéntico al calculado anteriormente?*

DESCRIPCIÓN DE LA INFORMACIÓN

Tal como lo hemos apuntado en el capítulo 1, el proceso de extracción de información consiste, en la mayoría de los casos, en la obtención de una muestra aleatoria de una población grande y, una vez obtenida la muestra, se procede al estudio de la información que ella contiene. El estudio de la muestra comienza con la que se denomina **descripción de la información** la cual consiste, a su vez, en la *presentación*, *organización* y *resumen* de los datos de la muestra.

Ordenamiento, clasificación y presentación de los datos

La primera forma con la que usualmente el analista se encuentra, es una tabla de datos **crudos**, es decir, los datos dispuestos de la manera en que los tomó el operador. Comúnmente, los datos se registran en cuadros, tablas o planillas. Por ejemplo, el Cuadro 1 contiene 100 datos correspondientes a las mediciones de diámetros de espigas de maíz en milímetros tal como fueron registrados por el técnico en el campo experimental, es decir, en el orden en que fueron leídos. Viendo el Cuadro 2.1, se podrían detectar algunas características aisladas de los números tales como números muy grandes o muy pequeños en comparación con los demás o, quizás, algún vacío de valores en algún segmento del Cuadro, pero no mucho más que eso. Para poder extraer más información de los datos, éstos deben estar **clasificados** u **organizados**.

Cuadro 2.1. Cien mediciones de diámetros de espigas de maíz, en milímetros.

56.0	51.8	54.4	53.0	54.3
41.0	51.0	51.8	54.4	52.5
53.1	46.1	44.9	49.0	53.8
46.0	45.6	58.0	55.4	53.7
40.2	45.2	52.3	55.4	54.6
53.8	51.1	49.0	65.2	59.6
47.7	48.3	51.0	63.8	60.0
51.6	47.6	53.3	59.1	55.3
44.4	51.2	60.7	52.6	39.7
52.7	50.1	54.7	61.0	43.0
44.6	46.4	56.5	53.0	42.0
51.5	40.0	52.7	51.4	39.7
47.2	55.1	55.5	61.0	44.6
47.5	52.5	52.3	57.2	42.6
44.0	51.1	50.0	55.3	43.0
50.0	51.7	49.5	56.3	39.0
48.4	54.3	52.0	58.7	46.9
54.0	50.6	53.5	51.4	41.6
46.0	46.7	55.0	64.6	43.3
51.3	47.7	43.0	54.2	46.7
Total: 5093.1				

Cuadro 2.2. Datos del Cuadro 2.1., clasificados en orden ascendente.

39.0	46.0	50.6	52.7	55.3
39.7	46.0	51.0	52.7	55.3
39.7	46.1	51.0	53.0	55.4
40.0	46.4	51.1	53.0	55.4
40.2	46.7	51.1	53.1	55.5
41.0	46.7	51.2	53.3	56.0
41.6	46.9	51.3	53.5	56.3
42.0	47.2	51.4	53.7	56.5
42.6	47.5	51.4	53.8	57.2
43.0	47.6	51.5	53.8	58.0
43.0	47.7	51.6	54.0	58.7
43.0	47.7	51.7	54.2	59.1
43.3	48.3	51.8	54.3	59.6
44.0	48.4	51.8	54.3	60.0
44.4	49.0	52.0	54.4	60.7
44.6	49.0	52.3	54.4	61.0
44.6	49.5	52.3	54.6	61.0
44.9	50.0	52.5	54.7	63.8
45.2	50.0	52.5	55.0	64.6
45.6	50.1	52.6	55.1	65.2

Una forma muy simple de organizar la información contenida en los datos consiste en disponerlos en orden a su magnitud, es decir, clasificarlos en orden ascendente o descendente. En el Cuadro 2.2 se han dispuesto las 100 mediciones del Cuadro 2.1 en orden ascendente.

Con los datos clasificados como en el Cuadro 2.2 se pueden hacer algunas cosas más que con los datos crudos como, por ejemplo, detectar cuáles son los valores máximo y mínimo del conjunto o ver si hay alguna discontinuidad en la secuencia de los números, o ver si los números tienen alguna tendencia a agruparse en alguna zona determinada. Pero, aún con las ventajas que presenta, en la mayoría de los casos la clasificación no le basta al investigador o al ingeniero para alcanzar sus objetivos. Un paso más decisivo en ese sentido lo representa la **condensación** de los datos en una **tabla** o **distribución de frecuencias.** En el Cuadro 2.3 se presenta la tabla de frecuencia correspondiente a los datos de los Cuadros 2.1 y 2.2.

Cuadro 2.3. Tabla de frecuencias correspondiente a los datos de los cuadros 2.1 y 2.2.

Clase	Punto medio (m_i)	Frecuencia de clase (f_i)
(35 – 40]	37.5	4
(40 – 45]	42.5	14
(45 – 50]	47.5	21
(50 – 55]	52.5	40
(55 – 60]	57.5	15
(60 – 65]	62.5	5
(65 – 70)	67.5	1
	Total	**100**

La tabla de frecuencias consiste en el agrupamiento de la masa de datos clasificados en un número reducido de grupos o clases delimitados por valores preestablecidos (**intervalos de clase**). Ya no existen más los valores individuales de los datos. Esta reducción implica, ciertamente, un cierto grado de pérdida de información porque, por ejemplo, del dato correspondiente al diámetro de 46.9 sólo sabemos ahora que está en algún lugar dentro de la tercera clase de la tabla de frecuencias. En este curso solo veremos distribuciones de frecuencia con intervalos de clase uniformes, es decir que las tablas de frecuencia tendrán todos los intervalos con el mismo ancho. En la distribución del Cuadro 2.3, todos los intervalos tienen un ancho igual a 5 milímetros.

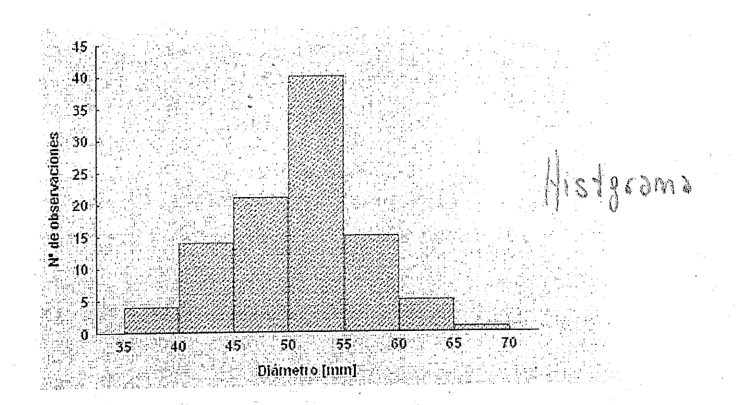
Los valores extremos de cada intervalo de clase son los **límites inferior** y **superior** del intervalo. Por ejemplo, el límite inferior de la cuarta clase de la tabla es 50 mm y el límite superior de la misma es 55 mm. Debemos notar que el valor del límite superior de una clase puede coincidir con el del límite inferior de la siguiente pero el dato correspondiente a ese valor debe pertenecer a una y solo una de las clases. Para eludir esta ambigüedad se utiliza el símbolo "]" para indicar la inclusión y el símbolo "(" para indicar la exclusión de ese valor. Por ejemplo, el dato 50 pertenece a la 3ª clase y no a la 4ª. En la tercera columna de la tabla se escriben **las frecuencias absolutas** correspondientes a cada clase. Las frecuencias absolutas no son otra cosa que la cantidad de datos que hay en cada clase. Por ejemplo, hay 40 datos dentro de la 4ª clase y 15 datos dentro de la 5ª. La suma de las frecuencias de clase (f_i) debe ser, obviamente, igual al total de datos en la muestra (usaremos el símbolo n para denotar el número de datos cuando se trate de una muestra y el símbolo N, cuando se trate de una población). Otro punto importante de cada clase es la **marca de clase** que no es otra cosa que el punto medio entre ambos límites. Por ejemplo, la marca de la 6ª clase es 62.5. En cuanto al número de clases a emplear para construir la tabla de frecuencias, eso depende de varias consideraciones pero, a modo de regla empírica, digamos que el número de clases debería estar entre 5 y 15. En nuestro ejemplo hay 100 datos y la amplitud total (es decir, la diferencia entre el máximo y el mínimo) es de 65.2 – 39.0 = 26.2. Para estos datos se eligió un ancho para los intervalos de clases de 5 mm y, por tanto, un total de 7 clases.

La tabla de frecuencias, a pesar de la reducción en la información que implica, presenta una serie de ventajas. Por ejemplo, utilizando tablas de frecuencias es más fácil comparar dos conjuntos de datos. Además, es más fácil obtener las medidas que permiten resumir la información en unos pocos números. Finalmente, la tabla de frecuencias hace mucho más fácil la obtención de gráficos representativos de la distribución de los datos en la muestra o en la población.

Tablas y representaciones gráficas de frecuencias

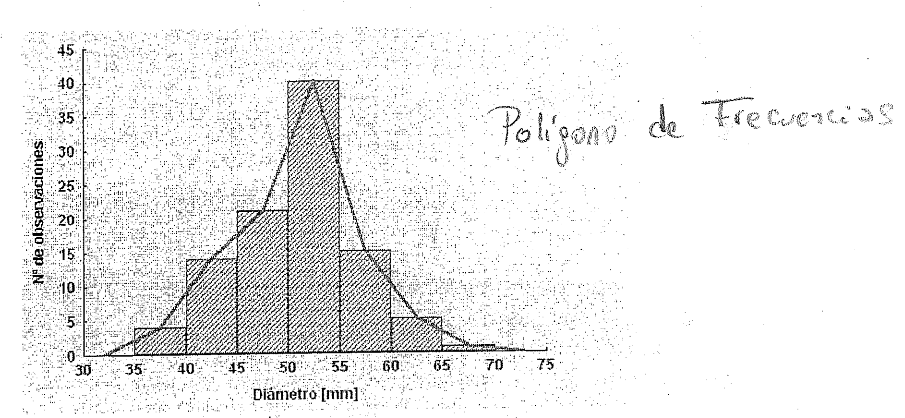
Existen muchas maneras de representar gráficamente una distribución de frecuencias. En este curso veremos tres de las más importantes: el **histograma**, el **polígono de frecuencias** y el **diagrama de caja y bigotes**. Un histograma es una representación en la cual se inscriben en el eje de abscisas los valores de la variable en estudio y en el eje de ordenadas los valores de las frecuencias. El histograma correspondiente a los datos del Cuadro 2.3 se presenta en la figura siguiente:

Figura 2.1. Histograma correspondiente a los datos del cuadro 1.3.



El polígono de frecuencias se obtiene, simplemente, uniendo mediante una línea poligonal los puntos medios en la cima de la barras del histograma de la distribución. En la siguiente figura se representan conjuntamente el polígono de frecuencia y el histograma correspondientes a los datos del cuadro 2.3:

Figura 2.2. Polígono de frecuencia e histograma correspondientes a las datos del cuadro 2.3.



El histograma es una representación muy completa de la distribución de frecuencias y superior al polígono pero, con todo, el polígono tiene utilidad en muchas instancias. Por ejemplo, el polígono es especialmente útil cuando se desean comparar dos distribuciones puesto que la superposición de los histogramas daría un gráfico confuso y difícil de interpretar mientras que la superposición de los polígonos deja espacio para una lectura cómoda y rápida. Otra ventaja del polígono de frecuencias es que puede, en ocasiones, ayudar a descubrir si hay alguna función matemática que pueda describir eficazmente la distribución real subyacente a la totalidad de los datos de la población.

De los polígonos de frecuencias a las curvas poblacionales

Si tomásemos una muestra muy grande podríamos acercarnos a la verdadera distribución de frecuencias de la población y cuanto más grande sea la muestra más cerca de aquella estaremos. Pero raras veces se puede tomar una muestra tan grande que pueda absorber las irregularidades causadas por el tamaño de las muestras pequeñas. En cambio, una muestra pequeña eficientemente tomada puede ser muy útil en sugerir la verdadera curva de la población (**curva poblacional**) mediante una función matemática derivada de los datos. Los tipos de curva poblacional más comunes se presentan en la siguiente figura:

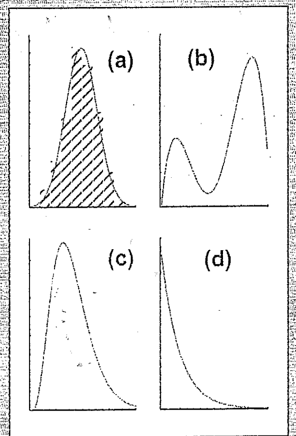


Figura 2.3. Ejemplos más frecuentes de curvas poblacionales:
(a) Distribución normal.
(b) Distribución bimodal.
(c) Curva asimétrica positiva.
(d) Curva de J invertida

Poder contar con una descripción matemática ajustada de la distribución de una variable en la población puede ser muy útil en el momento de tomar decisiones a partir de datos muestrales. Algunas clases más adelante haremos suposiciones acerca de las distribuciones de las variables en las poblaciones basándonos en distribuciones aproximadas desde las muestras.
Los tipos de curva poblacional más comunes se representan en la figura con las letras (a), (b), (c) y (d). La curva (a) representa una distribución de frecuencias muy común de hallar en la naturaleza y en los problemas de ingeniería y que describiremos en un capítulo posterior: la distribución **normal**. La curva (b) es una distribución **bimodal**, es decir, una distribución que presenta dos puntos de máxima frecuencia denominados **modas**. Definiremos a la **moda** algunas páginas más adelante. Las curvas bimodales suelen observarse en poblaciones que esconden dos distribuciones internas. La curva del tipo (c) es una curva **asimétrica positiva**, es decir, una curva asimétrica con su cola más larga hacia la derecha de los valores del eje x. Finalmente, la curva (d) es una curva en forma de **J invertida**

El tercer tipo de representación que veremos es el **diagrama de caja y bigotes**. Este tipo de gráfico presenta los valores de la variable en el eje de ordenadas, contrariamente al histograma y al polígono que presentan los valores de la variable sobre el eje de abscisas. Consiste en una caja que representa el 50% central de la distribución de los datos ordenados, es decir, desde el dato que deja por detrás suyo (en orden ascendente) al 25% de los datos, hasta el dato que deja por detrás suyo (en orden ascendente) al 75% de los datos. Mediante los bigotes pueden representarse diferentes medidas aunque lo más común es que se represente a los valores máximo y mínimo de la distribución. Finalmente, mediante un símbolo especial (una estrella, un segmento, un cuadrado, etc.) se representa la **mediana** de la distribución, es decir, el valor que tiene por debajo suyo al menos el 50% de los datos y por encima al menos el otro 50%. Definiremos más delante a la mediana. En la siguiente figura se presenta el diagrama de caja y bigotes de la distribución de frecuencias del Cuadro 2.3.

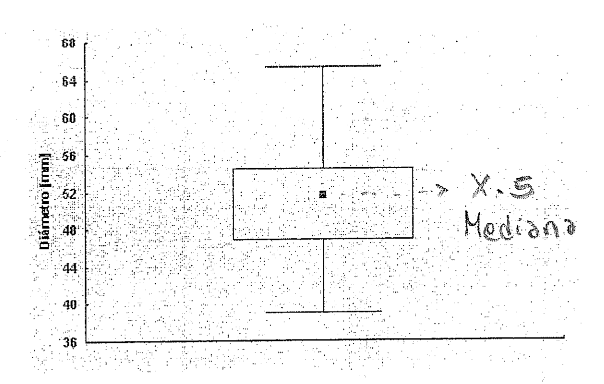


Figura 2.4. Diagrama de caja y bigotes de la distribución de frecuencias del cuadro 2.3.

Variables Cuantitativas

Frecuencias relativas

Las frecuencias relativas se obtienen a partir de las frecuencias absolutas de una manera muy simple: se divide cada frecuencia absoluta por el número total de datos de la muestra (o de la población), es decir, f_i/n ó f_i/N y se las denota como f_{ri}. En el siguiente cuadro se presenta la distribución de frecuencias relativas correspondiente a los datos de diámetro.

Cuadro 2.4. Frecuencias relativas correspondientes a los datos de diámetro de espigas de maíz. (Muestra 1)

Clase	Frecuencia (f_i)	Frecuencia relativa (f_{ri})
(35 – 40]	4	0.040
(40 – 45]	14	0.140
(45 – 50]	21	0.210
(50 – 55]	40	0.400
(55 – 60]	15	0.150
(60 – 65]	5	0.050
(65 – 70]	1	0.010
Total	**100**	**1.000**

Una de las grandes utilidades de la distribución de frecuencias relativas es que permite comparar distribuciones de frecuencias correspondientes a datos de diferente magnitud. Veremos un ejemplo de su utilidad. Supongamos que queremos comparar nuestra distribución de frecuencias de diámetros de espigas de maíz con otra distribución también de diámetros de espigas de maíz pero correspondiente a una muestra más grande de n = 200.

Los datos correspondientes a la primera muestra (n = 100) se presentan en el cuadro 2.4. Los datos correspondientes a la segunda muestra (n = 200) con la cual se desea comparar la primera, se presentan en Cuadro 2.5.

Cuadro 2.5. Frecuencias relativas correspondientes a los datos de diámetro de espigas de maíz. (Muestra 1)

Clase	Frecuencia (f_i)	Frecuencia relativa (f_i)
(35 – 40]	5	0.025
(40 – 45]	10	0.050
(45 – 50]	37	0.185
(50 – 55]	70	0.350
(55 – 60]	40	0.200
(60 – 65]	29	0.145
(65 – 70]	9	0.045
Total	**200**	**1.000**

Los polígonos de frecuencias de ambas distribuciones permitirán observar cuál es la utilidad del cálculo de las frecuencias relativas. En la figura 2.5. se presentan las frecuencias *absolutas* de ambas distribuciones y, como puede verse claramente, ambos polígonos no se pueden comparar, simplemente, porque la segunda muestra es más grande que la primera y, por esta razón, el polígono de frecuencias absolutas refleja este hecho.

En cambio, si graficamos los polígonos de frecuencias relativas de ambas muestras la comparación resulta válida y pueden verse las diferencias entre ambas muestras sobre una base homogénea. En la figura 2.6. pueden verse los polígonos de las frecuencias relativas de ambas muestras y se nota claramente como, por ejemplo, en la muestra 2 las espigas con diámetros superiores son un poco más frecuentes, en términos relativos, que en la muestra 1 mientras que las espigas con diámetros bajos son menos frecuentes, en general, en la muestra 1 que en la muestra 2.

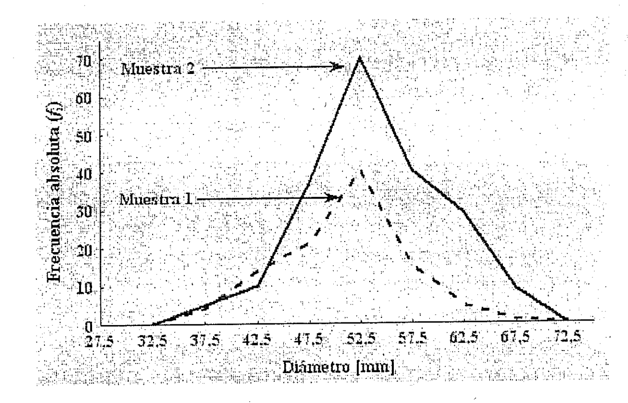


Figura 2.5. Polígonos de frecuencias absolutas de la muestra 1 y la muestra 2.

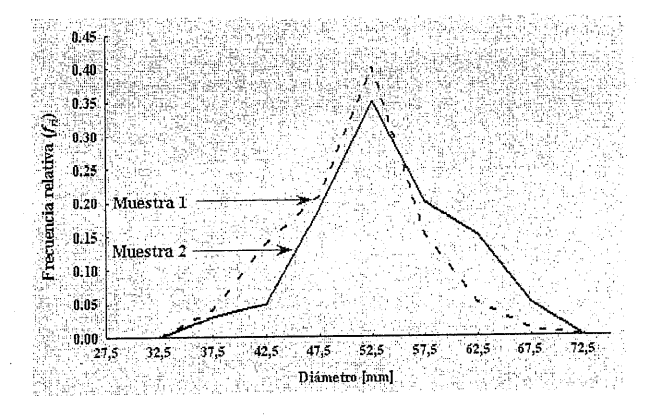


Figura 2.6. Polígonos de frecuencias relativas de las muestras 1 y 2.

Frecuencias acumuladas

Muchas veces, el interés del investigador no está puesto en la frecuencia absoluta o relativa de un determinado valor o intervalo de clase sino en el conjunto de valores que está por encima o por debajo de un valor específico. Por ejemplo, el *número o porcentaje de animales de un rodeo que pesa, por lo menos, 350 kg*, o el *número de plantas de trigo que presentan, a lo sumo, dos espigas infectadas por un hongo patógeno*. Para poder contestar rápidamente este tipo de preguntas se calculan las denominadas **frecuencias acumuladas**, tanto absolutas como relativas.

El cálculo de las frecuencias acumuladas (F_i o F_i/n) se puede hacer en forma ascendente o descendente y es muy simple: en el primer caso, consiste en acumular los valores de las frecuencias absolutas (o relativas) hasta alcanzar el máximo valor, n (o 1); en segundo, consiste en ir restando o desacumulando las frecuencias absolutas (o relativas) desde el máximo valor, n (o 1) hasta llegar a 0. Ahora se aplicarán estos cálculos al ejemplo de los diámetros de las espigas de maíz con el que se viene trabajando.

Cuadro 2.6. Cálculo de frecuencias acumuladas.

		Creciente		Decreciente	
Clase	f_i	F_i	F_i/n	F_i	F_i/n
(35 – 40]	4	4	0.04	100	1.00
(40 – 45]	14	18	0.18	96	0.96
(45 – 50]	21	39	0.39	82	0.82
(50 – 55]	40	79	0.79	61	0.61
(55 – 60]	15	94	0.94	21	0.21
(60 – 65]	5	99	0.99	6	0.06
(65 – 70]	1	100	1.00	1	0.01

Las frecuencias acumuladas se representan mediante el gráfico como el de la figura 2.7.:

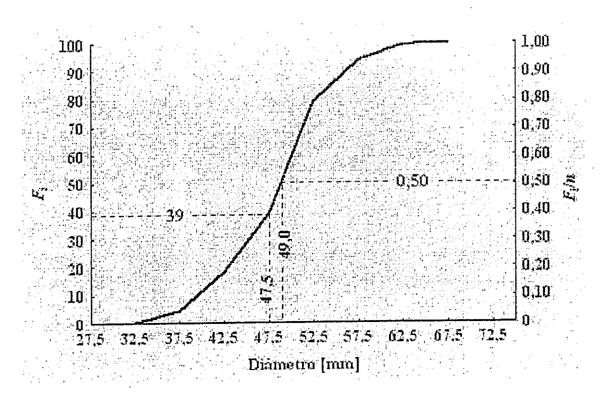


Figura 2.7. Ojiva, representación de frecuencias acumuladas.

La distribución de frecuencias acumuladas se utiliza para calcular gráficamente valores tanto sobre el eje de abscisas como sobre el eje de ordenadas. En la figura anterior, sobre el eje de ordenadas de la izquierda, se representaron las frecuencias absolutas acumuladas y sobre el eje de ordenadas de la derecha, las frecuencias relativas acumuladas. Por ejemplo, como se muestra en la figura, si queremos conocer la frecuencia acumulada absoluta correspondiente a los 47.5 mm de diámetro, solo tenemos que ascender en línea recta desde la posición 47.5 sobre el eje de abscisas hasta llegar a la ojiva y, a partir de ella, seguir en línea recta horizontal hasta interceptar el eje de ordenadas de la izquierda, para obtener el valor 39. También podemos usar la ojiva en forma inversa. Por ejemplo, si queremos saber cuál es el valor que acumula el 50% de la observaciones de diámetro, partimos del punto 0.50 sobre el eje de ordenadas de la derecha, seguimos en línea recta horizontal hasta llegar a la ojiva y, desde

allí, descendemos en línea recta vertical hasta llegar al eje de abscisas, para obtener un valor aproximado de 49 mm.

Variables cualitativas

Hasta ahora hemos visto tablas de frecuencias y representaciones gráficas para variables cuantitativas pero todo esto también puede hacerse para variables cualitativas. La tabla de frecuencias correspondiente a una variable cualitativa muestra, simplemente, las frecuencias, tanto absolutas como relativas, tanto simples como acumuladas, para cada una de las categorías en las que está clasificada la variable. Supongamos, por ejemplo, que se recibe una encomienda de 200 unidades de un material clasificado según su grado de pureza en 5 categorías: *muy puro*, *puro*, *mediano*, *impuro* y *muy impuro*. Una vez hecho el recuento se obtienen los resultados de la siguiente tabla:

Cuadro 2.7.

	Frecuencia	
	Absoluta	Relativa
Muy puro	35	0.175
Puro	59	0.295
Mediano	52	0.260
Impuro	42	0.210
Muy impuro	12	0.060
Total	**200**	**1.000**

La representación gráfica que puede usarse es un diagrama de barras verticales u horizontales en el que se indican las categorías de la variable sobre el eje de abscisas y las frecuencias, sobre el de ordenadas. A continuación se presenta un diagrama de barras verticales correspondiente al ejemplo de la pureza de los materiales. Cabe aclarar que los anchos de las barras son enteramente arbitrarios y no tienen significado práctico aunque deben ser iguales entre sí.

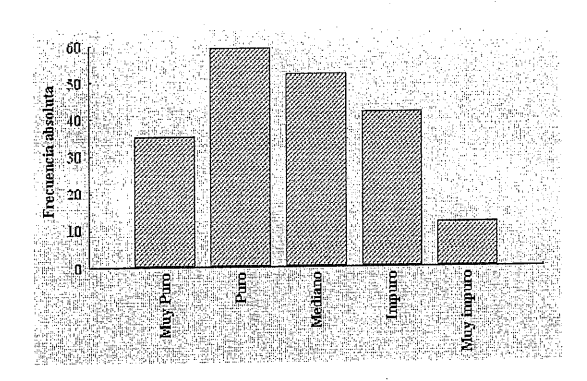


Figura 2.8. Diagrama de barras verticales.

Medidas resumen de la información

El proceso de resumen de la información no se detiene con la distribución de frecuencias. Aún se puede resumir mucho más sin que por eso se pierda la posibilidad de rescatar la información verdaderamente útil y que resulta de interés. El proceso continúa con la obtención de las denominadas **medidas resumen**. Veremos dos clases básicas de medidas: (a) las medidas de **posición** (también llamadas medidas de **tendencia central**) y, (b) las medidas de **dispersión**.

Medidas de posición

Las medidas de posición o de tendencia central dan una idea de cómo es la estructura de los datos, especialmente, la región central de la distribución de los mismos y, por ese motivo, reciben la denominación general de **promedios**. Aunque no siempre, algunas medidas de posición no están relacionadas con la región central de la distribución sino con otras partes de la misma. Las medidas promedio guardan cierta semejanza con el concepto de centro de gravedad de un cuerpo físico. Hay muchas medidas de posición pero en este curso veremos solamente tres: (i) los **cuantiles** y la **mediana**, (ii) la **moda** y, (iii) la **media aritmética**.

Los cuantiles

Los cuantiles son medidas que se obtienen sobre la distribución de los datos clasificados. Una vez ordenados los datos en orden ascendente, se buscan en los mismos, ciertas posiciones específicas de interés. Las tres clases de cuantiles más comunes son: (i) los **cuartiles**, (ii) los **deciles** y, (iii) los **percentiles**. Los cuartiles son posiciones que dividen la distribución de los datos en cuatro secciones. La primera va desde el valor mínimo hasta el valor que deja por debajo suyo, por lo menos, al 25% de los datos y por encima suyo, por lo menos, al 75% de los mismos; este valor recibe el nombre de **primer cuartil** y se lo simboliza q_1. La segunda va desde el primer cuartil hasta el valor que deja por debajo suyo, por lo menos, al 50% de los datos y por encima suyo, por lo menos, al otro 50% de los mismos; este valor recibe el nombre de **segundo cuartil** o **mediana** de la distribución y se lo simboliza q_2, o $x_{.5}$. La tercera va desde la mediana hasta el valor que deja por debajo suyo, por lo menos, al 75% de los datos y por encima suyo, por lo menos, al 25% de los mismos; éste valor recibe el nombre de **tercer cuartil** y se lo simboliza q_3. Y la última que va desde q_3 hasta el valor máximo.

La moda

La **moda** simbolizado x_m, es, simplemente, el valor más frecuente de la distribución. Dada su definición, es posible encontrarse con distribuciones cuyos valores tengan, todos, la misma frecuencia: en ese caso, la distribución de los datos carece de moda. O podría darse el caso de una distribución que posea más de una moda. Por ejemplo, en la sección sobre curvas poblacionales, vimos una curva que poseía dos modas (curva bimodal).

La media aritmética

La **media aritmética** es, simplemente, el resultado de dividir la suma de todos los valores por n, el tamaño de la muestra (o N, si se tratara de una población) y se la simboliza $\overline{x}$:

$$\begin{aligned}\overline{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i\end{aligned} \tag{2.1}$$

Las calculadoras de bolsillos con modo estadístico (SD) permiten calcular la media aritmética (aparecen con el símbolo $\overline{x}$). Aplicando la fórmula a los datos de muestra del cuadro 2.1, obtenemos:

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{100}\cdot(5093.1) \\ &= 50.931\,mm\end{aligned}$$

La media aritmética tiene las siguientes dos propiedades de gran interés para el análisis de datos:

I. que la suma de los desvíos de todos los valores de la muestra con respecto a la media aritmética es igual a 0:

$$\sum_i (x_i - \bar{x}) = 0$$

II. que la suma de las desviaciones de los datos con respecto a la media elevadas al cuadrado, es menor que la suma de las desviaciones de los datos con respecto a cualquier otro valor elevadas al cuadrado.

$$\sum_i (x_i - \bar{x})^2 = mín.$$

La última propiedad cobrará relevancia cuando se definan las medidas de dispersión.

Cuando la muestra presenta valores repetidos muchas veces, conviene utilizar la **media aritmética ponderada** que se calcula con la fórmula general que se dio más arriba salvo que se indica mediante factores (**ponderaciones**) la cantidad de veces que se repite cada valor. Por ejemplo, supongamos la siguiente muestra: 1, 1, 3, 3, 3, 3, 3, 4, 5, 5, 5, 5, 5, 5, 5, 6, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 8, 9 ,9, 9, 9, 9, 9, 9, 13, 15, 15, 15, 15, 15, 18, 23, 23, 24, 24, 24, 24, 24, 25, 25. Son 50 datos, algunos de los cuales se repite varias veces. Entonces, en lugar de calcular la media con la fórmula anterior, sumando los valores uno por uno, se multiplica cada valor por su ponderación y se divide el total por n (en este ejemplo, n = 50):

$$\bar{x} = \frac{1\cdot 2+3\cdot 5+4\cdot 1+5\cdot 7+6\cdot 1++7\cdot 3+\cdots+24\cdot 5+25\cdot 2}{2+5+1+7+1+3+\cdots+5+2} = \frac{532}{50} = 10.64$$

Una fórmula general para este cálculo es: $\bar{x}_w = \frac{1}{\sum_i w_i}\cdot\left(\sum_i x_i \cdot w_i\right)$ (2.2)

donde $\bar{x}_w$ es la media aritmética ponderada, x_i son los valores de las observaciones individuales y w_i son las ponderaciones

Medidas de dispersión

Las medidas de posición, especialmente los promedios (media, mediana y moda), como se dijo antes, dan una idea de cuál es el "centro de gravedad" de la masa de datos pero nada dicen de cómo están distribuidos los datos alrededor de esos puntos centrales. Por ejemplo, la distribución formada por los números 1, 4, 8, 13, 18, 22 y 25 y la distribución formada por los números 10, 11, 12, 13, 14, 15 y 16 tienen, ambas, la misma media aritmética, $\bar{x}$ = 13 pero no cabe ninguna duda de que la primera de las distribuciones tiene los datos más dispersos alrededor del punto central, que la segunda. Entonces, para completar la caracterización de una distribución de frecuencias, se necesita contar con alguna medida de esa dispersión. En este curso veremos tres principales, la **amplitud**, la **amplitud intercuartil** y la **variancia** y otras dos que derivan de la variancia: el **desvío standard** y el **coeficiente de variación**.

Amplitud

La amplitud es la medida de dispersión más simple. Esta medida también se la conoce con el nombre de rango, aunque es más apropiado el término amplitud. En un conjunto de n observaciones x_1, x_2, x_3, ...x_n la amplitud se define como la diferencia entre el máximo (x_{max}) y el mínimo (x_{min}). A pesar de la facilidad de cálculo y la simpleza de esta medida, la amplitud puede resultar insensible a la variación de los datos, sobretodo en conjuntos grandes de datos.

Amplitud intercuartil

La amplitud intercuartil, como su nombre lo indica claramente, es la diferencia, en valor absoluto, entre q_1 y q_3 e incluye, por esta misma razón, el 50% central de la distribución de frecuencias. Es la que determina la longitud de la caja en el diagrama de caja y bigotes que vimos páginas atrás.

Variancia y desvío standard

La **variancia** es una medida refleja la dispersión de los datos alrededor se define como el promedio de los cuadrados de los desvíos de los datos con respecto a su media. Para el caso de un conjunto de n datos de la variable x (x_1, x_2, ..., x_n), la varianza se calcula como:

$$Variancia(X) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (2.3)$$

Por razones que detallaremos en el capítulo 6, cuando se quiere estimar la varianza de una variable aleatoria en una población a partir de los datos de una muestra aleatoria se utiliza la fórmula 2.4.

$$s_{n-1}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (2.4)$$

Claramente, s_{n-1}^2 no es el promedio de los cuadrados de los desvíos. Sin embargo, como se usa para calcular el valor estimado de la variancia en la población a partir de los datos de la muestra, esta variable es llamada comunmente *variancia muestral*.

Como la variancia es un promedio de desvíos elevados al cuadrado, sus unidades son las unidades originales elevadas al cuadrado. Para eliminar esta inconveniencia, se suele medir la dispersión de los datos por medio del **desvío standard** que no es otra cosa que la raíz cuadrada de la variancia. Las calculadoras de bolsillos con modo estadístico (SD) permiten calcular tanto s_n como s_{n-1}.

Ilustraremos el cálculo de la variancia con un ejemplo. En el siguiente cuadro se presentan los registros de los rendimientos en grano de un híbrido de girasol (en Kg/parcela) en 10 ensayos experimentales:

125	120	118	133	127	119	130	124	131	121

Para aplicar la fórmula, primeramente debemos calcular la media aritmética la cual resulta ser $\bar{x} = \frac{1248}{10} = 124.8$. Luego, podemos ordenar las cifras en forma de cuadro para facilitar los cálculos (x representa el rendimiento, en Kg/parcela):

X	120	125	118	133	127	119	130	124	131	121	**1248**
$(x_i - \bar{x})^2$	23,04	0,04	46,24	67,24	4,84	33,64	27,04	0,64	38,44	14,44	**255,6**

Luego:

$$s^2_{(n)X} = \frac{255.6}{10} = 25.56 \Rightarrow s_{(n)X} = \sqrt{25.56} \cong 5.055 \text{ Kg/parcela;}$$

y:

$$s^2_{(n-1)X} = \frac{255.6}{(10-1)} \cong 28.4 \Rightarrow s_{(n-1)X} = \sqrt{28.4} \cong 5.329 \text{ Kg/parcela.}$$

La muestra ha sido pequeña y, por esta razón, hay una diferencia más o menos notoria entre ambas fórmulas de variancia pero con muestras más o menos grandes la diferencia se hace insignificante.

Coeficiente de Variación

Cuando se necesita comparar el grado de variabilidad en la información entre dos muestras correspondientes a poblaciones diferentes en la magnitud de los datos, el solo uso del desvío standard no es suficiente porque surgirán diferencias que se deben a la naturaleza de los datos y no a las variaciones de las muestras en sí. En ese caso, se recurre a una medida relativa de la variabilidad denominada **coeficiente de variación** (*cv*) que es, simplemente, el cociente entre el desvío standard y la media aritmética, multiplicado por 100. Para s_{n-1} tenemos:

$$cv = \frac{s_{n-1}}{\bar{x}} \cdot 100 \qquad (2.5)$$

Ejemplo.
Se cuenta con una muestra de pesos de cerdos y con otra muestra de pesos de gallinas, y se desea saber cuál es comparativamente más variable. Los valores de medias y desvíos son los siguientes:

Cerdos: $\bar{x}_1 = 324$ Kg.; $s_{(n-1)1} = 38.8$ Kg.;

Gallinas: $\bar{x}_2 = 1.6$ Kg.; $s_{(n-1)2} = 0.299$ Kg.

Obviamente, los pesos de los cerdos tienen una variabilidad absoluta mucho mayor pero, ¿son realmente, más variables en relación con su media? Calculemos los respectivos *cv*:

$$cv(x_1) = \frac{s_{(n-1)1}}{\bar{x}_1} = \frac{38.8}{324} \cdot 100 = 12 \quad \text{y}$$

$$cv(x_2) = \frac{s_{(n-1)2}}{x_2} = \frac{0.299}{1.600} = 18.7$$

Vemos que la variabilidad relativa en las gallinas es un 57% mayor que la correspondiente a los cerdos, aunque su desvío standard sea menor.

Cálculos de media y variancia partiendo de distribuciones de frecuencia (datos agrupados)

La media aritmética y la variancia suelen calcularse al mismo tiempo para datos agrupados, porque, para ambas medidas, puede usarse la misma hoja de trabajo. Adviértase que ambas medidas requieren todos los valores individuales de la muestra. Pero sabemos que esos valores se pierden en el proceso de organizar una distribución de frecuencias. Esta dificultad se evita si usamos el punto medio (m_i) de la *i*-ésima clase para representar todos y cada uno de los valores individuales de dicha clase.

Repitiendo este procedimiento para todas las clases se obtiene el valor total de toda la distribución. En consecuencia, la media aritmética para datos agrupados, con *k* clases, puede definirse como:

$$\bar{x} = \frac{f_1 \cdot m_1 + f_2 \cdot m_2 + \cdots + f_k \cdot m_k}{f_1 + f_2 + \cdots + f_k} = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot m_i \qquad (2.6)$$

En términos de datos agrupados, la variancia puede definirse como:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot (m_i - \bar{x})^2 \quad \text{y} \qquad (2.7)$$

$$s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i \cdot (m_i - \bar{x})^2 \qquad (2.8)$$

o, más sencillamente:

Para los datos del Cuadro 3:

m_i	37.5	42.5	47.5	52.5	57.5	62.5	67.5	Total
f_i	4	14	21	40	15	5	1	100

obtenemos:

$$\bar{x} = \frac{4 \cdot 37.5 + 14 \cdot 42.5 + \cdots + 1 \cdot 67.5}{4 + 14 + \cdots + 1} = \frac{5085}{100} = 50.85 \text{ mm.}$$

Para la variancia:

m_i	37.5	42.5	47.5	52.5	57.5	62.5	67.5
$m_i - \bar{x}$	-13.35	-8.35	-3.35	+1.65	+6.65	+11.65	+16.65
f_i	4	14	21	40	15	5	1

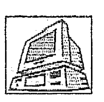
$$s_n^2 = \frac{(-13.35)^2 \cdot 4 + (-8.35)^2 \cdot 14 + \cdots + (+16.65)^2 \cdot 1}{100} = \frac{3652.75}{100} = 36.5275$$

$$\text{y} \quad s_{n-1}^2 = \frac{100}{100-1} \cdot 36.5275 = 36.8965 \text{.}$$

Las desviaciones standard: s_n = 6.044 mm y s_{n-1} = 6.074 mm.

Ejercicios

2.1 Como parte de un estudio de la producción vitivinícola de la provincia de La Rioja, se tomó una muestra aleatoria de 25 fincas del departamento de Chilecito y se registró su producción en tn/ha. Los datos obtenidos son figuran en la tabla:



Finca	Prod. (tn/ha)	Finca	Prod. (tn/ha)	Finca	Prod. (tn/ha)	Finca	Prod. (tn/ha)	Finca	Prod. (tn/ha)
1	13,8	6	16,4	11	15	16	13,2	21	14,4
2	14,6	7	15,8	12	14	17	14,7	22	11,9
3	16,8	8	12,6	13	13,8	18	17,6	23	16,3
4	14,6	9	17,3	14	14,2	19	14,7	24	13,5
5	16,1	10	14,5	15	13,5	20	14,3	25	15

a. *En este estudio ¿cuáles son las unidades muestrales, cuál es la muestra y cuál es la población representada por dicha muestra?* Producción

b. *¿Por qué es importante que la muestra sea tomada aleatoriamente? Imaginar una situación en la que esta condición no se cumpla, y explicar sus consecuencias.*

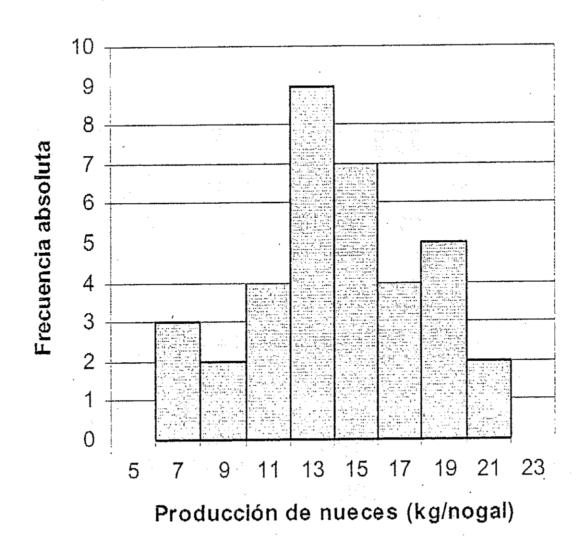
c. *Calcular la media aritmética y el desvío standard de los valores de producción registrados.* 14,244 8,2

d. *Construir una tabla de frecuencias relativas de los valores de producción registrados.*

e. *Construir un histograma de frecuencias relativas acumuladas.*

f. *¿En cuál clase se encuentra la mediana y en cuál se encuentra el tercer cuartil?*

2.2 El siguiente gráfico representa la distribución de frecuencias de la producción de nueces producidas por 36 nogales de una finca en la localidad de Tinogasta en la provincia de Catamarca.



a. *¿De qué tipo de gráfico se trata?*

b. *¿Cuál es el valor aproximado del tercer cuartil?*

c. *En otra finca en la provincia de San Juan se tomó una muestra similar y se encontró mayor producción por nogal y menor varianza*

entre nogales que en la de Tinogasta ¿Qué diferencias esperaría encontrar entre el gráfico correspondiente a los nogales de dicha finca y el gráfico que se presenta aquí?

2.3 Para evaluar la calidad del algodón proveniente de un establecimiento de la provincia de Catamarca, se obtuvo una muestra de 18 porciones de fibra tomadas al azar la salida de la desmotadora. A partir de cada porción, se obtuvo una medición independiente de la longitud de fibra. Los datos figuran en la tabla:

Unidad Muestral	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Longitud de fibra (mm)	28,9	34,1	30,6	31,1	35,8	29,5	32,9	36,2	32,0

Unidad Muestral	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*
Longitud de fibra (mm)	38,1	35,2	30,1	38,0	33,3	33,2	32,5	32,1	31,8

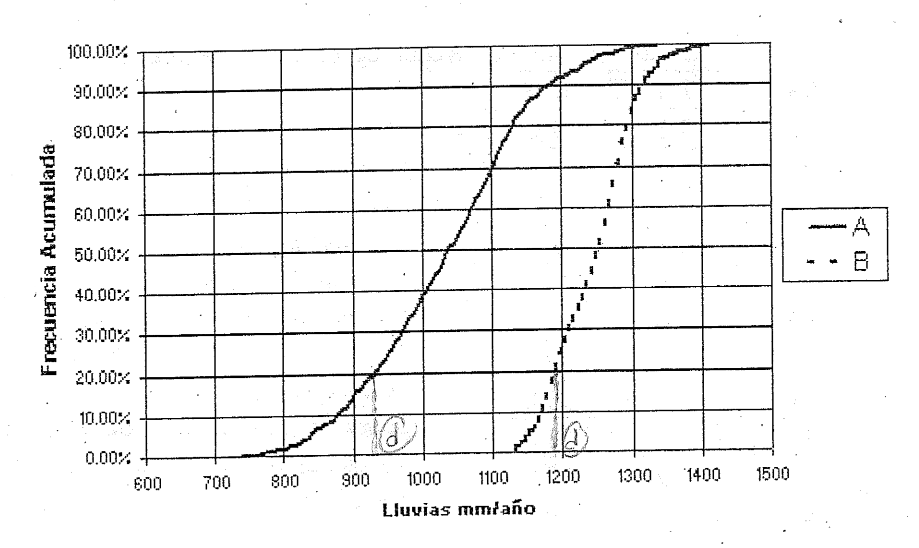
a. *Construir una tabla de frecuencias absolutas, relativas y relativas acumuladas.*

b. *Construir un gráfico de frecuencias relativas acumuladas y, sobre el mismo, identificar a la mediana de las longitudes de fibra*

c. *Calcular la media aritmética, la varianza y el desvío standard de las longitudes de fibra (mostrar los cálculos)*

d. *En un establecimiento del Chaco se analizó una muestra de 18 porciones de fibra y se encontró que la longitud media era de 2.98 cm y el desvío standard de 0.33 cm. ¿Cuál establecimiento produce fibras más largas y en cuál la longitud es más heterogénea?*

2.4 El siguiente gráfico de frecuencias relativas acumuladas ha sido construido con registros históricos de lluvias anuales de 2 localidades distintas (A y B):



a. *¿En cuál localidad llovió más en promedio?*

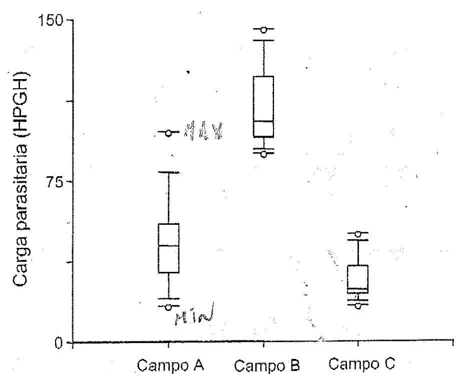
b. *¿En cuál localidad las lluvias fueron más variables entre años?*

c. *¿Cuál fue la lluvia anual mediana en la localidad B?*

d. *¿Qué valor de precipitación fue superado por el 80% de los años en cada localidad?*

e. *¿En cuál localidad fueron mas frecuentes los años con precipitaciones entre 900 y 1200 mm?* A

2.5 Los siguientes diagramas de caja representan la distribución de la carga parasitaria en ovinos de 30 días de edad en tres campos de la provincia de Corrientes. La carga parasitaria animal fue estimada a partir del recuento de huevos en las heces (número de huevos por gramo de heces HPGH).



a. *Diseñar un breve protocolo para obtener datos como los utilizados en este caso.*

b. *En este gráfico, ¿cuáles campos presentan distribuciones asimétricas? Justificar* B y C y A

c. *¿En cuál de los campos examinados se realiza aparentemente un mejor manejo sanitario? Discutir y justificar su respuesta.* C

d. *¿En cuáles campos la mediana del número de parásitos por animal no supera los 66 huevos por gramo de heces?* A y C

2.6 La tabla que se ve a continuación muestra los totales de precipitación (en mm) caídos en el mes de enero en las localidades de Esquel y Bariloche durante el periodo 1990-2001:

	Esquel	Bariloche
1990	7.9	12.3
1991	9.8	9.4
1992	3.2	7.5
1993	10.9	7.3
1994	5.7	9.3
1995	7.4	23.2
1996	8.4	27.5
1997	6.3	19.5
1998	5.1	16.2
1999	2.3	13.2
2000	6.9	19.8
2001	4.1	15
$\bar{x}$	6,5	15,92
DS	2,6	6,45
C	0,82	0

Según estos datos de precipitación:

a) *¿En cuál localidad ha llovido más en promedio en el mes de enero? Justificar la respuesta con un estadístico apropiado* Media

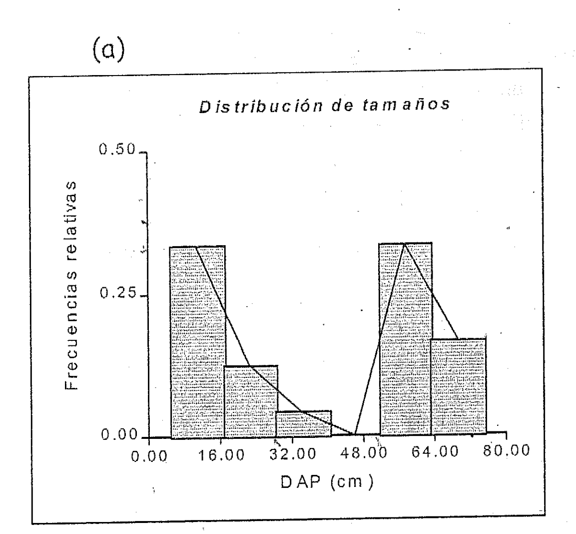
b) *¿En cuál localidad las precipitaciones han sido más variables? Justificar la respuesta con un estadístico apropiado*

c) *¿Cuál es la proporción de años en los cuales ha llovido menos de 8 mm en cada una de las localidades?*

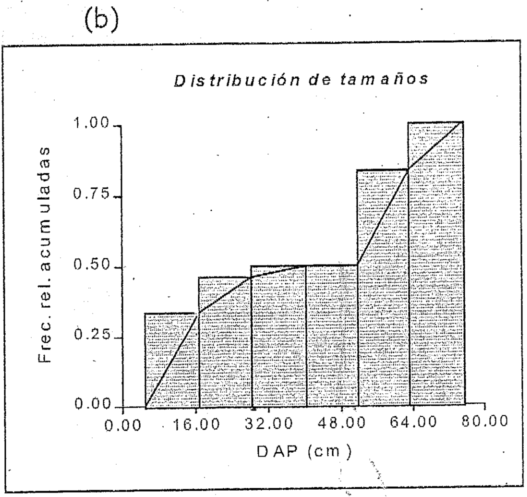
2.7 Las siguientes figuras representan la distribución de tamaños (DAP: diámetro a la altura del pecho) de una población de *Prosopis caldenia*

(caldén) localizada en Luan Toro, provincia de La Pampa. (a) histograma de **frecuencias relativas** y (b) histograma de **frecuencias relativas acumuladas**.

(a)



(b)



a. ¿La distribución de los tamaños es unimodal? Justifique. ¿Cuál es el valor aproximado de la o las modas?

b. ¿Qué porcentaje de fustes comercializables existe si el criterio es que superen los 48 cm de diámetro?

c. ¿Qué porcentaje de individuos no superan los 16 cm de diámetro?

d. ¿Qué porcentaje aproximado de individuos se hallan dentro del rango de 30 a 50 cm de diámetro?

CALCULO DE PROBABILIDADES

En el capítulo 2 hemos presentado formas para organizar, describir y presentar los datos de una variable aleatoria registrados en una muestra. El análisis de los datos muestrales tiene, en realidad, la finalidad la finalidad de conocer algo acerca de una población de la cual la muestra fue extraída. Utilizar información contenida en una muestra para extraer conclusiones acerca de la información desconocida contenida en una población implica un **riesgo** basado en la **incertidumbre** implícita en dicha **decisión**. La Estadística provee una manera racional de cuantificar y acotar tal incertidumbre y para ello utiliza una medida de la incertidumbre denominada **probabilidad**. La utilización del concepto de probabilidad y de los métodos para su cálculo constituye la base sobre la que se asienta la **toma de decisiones**. Como hemos dicho en una sección anterior, la toma de decisiones estará, generalmente, referida a la elección de un valor determinado para un parámetro desconocido o a la elección de algún conjunto de valores al cual se asume que dicho parámetro desconocido pertenece.

La existencia de incertidumbre acerca de un proceso físico implica la existencia de estados alternativos posibles para el mismo. Se cuenta con una determinada cantidad de información y se desea conocer una cantidad de información adicional, ordinariamente, la porción restante de la información total. Para ello, se debe contar con una enumeración del total de estados posibles del proceso. Además, se debe tener una medida de la posibilidad de ocurrencia para cada uno de dichos estados. Daremos, ahora, algunas definiciones.

Un **experimento aleatorio** es un proceso cuyos resultados no se conocen *a priori*. El conjunto de todos los resultados de un experimento aleatorio se denomina **espacio muestral** lo que denotaremos S. Cada uno de los resultados posibles contenidos en un espacio muestral es un **evento simple**. Dado que los estados posibles del proceso son alternativos, ellos no pueden ocurrir simultáneamente, por lo cual se los considera **mutuamente excluyentes**. Además, asumiremos que el espacio muestral contiene todos y cada uno de esos estados alternativos, por lo cual se dice que ese conjunto de eventos simples es **colectivamente exhaustivo**.

Ejemplos:

- Se arroja un dado de 6 caras, con una determinada cantidad de puntos en cada una de sus caras. Una cara contiene 1 punto, otra contiene 2 puntos, otra contiene 3 puntos, otra contiene 4 puntos, otra contiene 5 puntos y la última, 6 puntos. Por tanto, $S = \{1,2,3,4,5,6\}$.

- Se arroja al aire una moneda equilibrada. Entonces, $S = \{C,X\}$, donde C representa las caras y X, las cruces.

Un subconjunto de eventos simples del espacio muestral constituye un **evento compuesto** y se los suele denotar con letras mayúsculas.

Ejemplo:

- Cuando se arroja un dado, un evento compuesto es el subconjunto de las caras con un número impar de puntos: $A = \{1,3,5\}$. Otro evento compuesto es el subconjunto de las caras con una cantidad de puntos superior a 3: $B = \{4,5,6\}$.

Conjuntos

Repasaremos algunas operaciones básicas que se realizan entre conjuntos.

Unión de dos conjuntos

La unión de dos conjuntos A y B es el conjunto de elementos que pertenecen a *por lo menos uno* de los conjuntos A y B – es decir, a A o a B o a ambos. Simbolizamos esta operación como $A \cup B$ (se lee "A unión B").

Ejemplos

- Consideremos, por ejemplo $A = \{a,b,c,d,e,f\}$ y $B = \{a,e,i,o,u\}$, los cuales son subconjuntos del conjunto de todas las letras del alfabeto. El conjunto cuyos elementos están en A o en B o en ambos es el conjunto $A \cup B = \{a,b,c,d,e,f,i,o,u\}$.
- Si $A = \{x: x$ es un entero impar positivo$\}$ y $B = \{x: x$ es un entero impar positivo menor que $100\}$, entonces, $A \cup B = \{x: x$ es un entero impar positivo$\}$.

Intersección

La intersección de dos conjuntos A y B es el conjunto de elementos que pertenecen a A y a B. Simbolizamos esta operación como $A \cap B$ (se lee "A intersección B").

Ejemplo.

- Sean $A = \{1,2,3,4,5\}$ y $B = \{4,5,6,7,8\}$; entonces $A \cap B = \{4,5\}$

Complemento

Cuando hablamos de *complemento* de B, lo que denotamos B^c, entendemos el conjunto de todos los elementos del universo que no están en el conjunto B.

Ejemplo

- Sea $S = \{1,2,3,4,5,6,7,8,9,0\}$ y $U = \{1,3,5,7,9\}$. Luego, $U^c = \{2,4,6,8,0\}$.

Diferencia

La diferencia de dos conjuntos es el conjunto de todos los elementos del primer conjunto que no están en el segundo. $A - B = A \cap B^c$

Ejemplo

- Sean, otra vez, $A = \{1,2,3,4,5\}$ y $B = \{4,5,6,7,8\}$. Entonces, $A - B = \{1,2,3\}$ y $B - A = \{6,7,8\}$.

Aproximaciones a la medida de la probabilidad

Supuesto de eventos simples igualmente probables

Utilizando este supuesto, consideramos que todos los eventos incluidos en el espacio muestral tienen las misma chances de ocurrir. Por ejemplo, si arrojamos una moneda de masa homogénea, esperaremos que los dos eventos posibles, tendrán las mismas chances de ocurrir y dado que su suma debe ser igual a 1 (por el primer axioma), entonces, la probabilidad de cada evento será igual a 1/2, es decir que $P(X) = P(C) = 1/2$. Similarmente, si arrojamos el dado de 6 caras de masa homogénea que se mencionó en la página anterior, bajo el supuesto de eventos igualmente probables, $P(1) = P(2) = \cdots = P(6) = 1/6$. En general, cuando hay K resultados posibles igualmente probables, la probabilidad de cada uno de ellos será igual a $1/K$.

Entonces, si un espacio muestral tiene K resultados posibles y si un evento, A, que forma parte de ese espacio muestral contiene k eventos, la probabilidad de dicho evento es, simplemente, el cociente entre k y K:

$$P(A) = \frac{k}{K} \tag{3.1}$$

Ejemplos:

- Se extrae un naipe de una baraja inglesa de 52 cartas. Sea *A* el evento *corazón*. Entonces, teniendo en cuenta que hay 13 corazones en total en la baraja:

$$P(A) = \frac{13}{52} = \frac{1}{4}$$

- Otra vez, se extrae un naipe de una baraja inglesa de 52 cartas. Sea *B* el evento *número menor a 6*. Hay 5 cartas menores a 6 en cada palo, por lo cual, recordando que hay 4 palos, hay un total de 4 5 = 20 cartas cuyo número es menor a 6. Así que:

$$P(B) = \frac{20}{52} = \frac{5}{13}$$

- Finalmente, se extrae un naipe de la baraja inglesa de 52 cartas. Sea *A* el evento de que salga *un corazón o un trébol*. Hay 13 corazones y 13 tréboles en la baraja. Así que:

$$P(A) = \frac{26}{52} = \frac{1}{2} \quad \text{o}$$

$$P(C \cup T) = P(C) + P(T) = \frac{13}{52} + \frac{13}{52} = \frac{26}{52} = \frac{1}{2}$$

Frecuencia relativa de un evento

Cuando no es posible sostener el supuesto de eventos igualmente probables es necesario recurrir a otra manera de calcular las probabilidades. Volviendo al ejemplo de la moneda, si el supuesto de masa homogénea no se puede sostener, ¿qué probabilidades habrá que asignarles a los eventos *C* y *X*? Para poder responder a esa pregunta se recurre al cálculo de las frecuencias relativas de cada evento mediante experimentos repetidos. Se lanza repetidamente la moneda en cuestión y se registra la cantidad de cruces que salen. Y esa cantidad, en relación al total de veces que se lanzó la moneda, se toma como la probabilidad de que salga una cruz en el futuro.

Si se lanza la moneda, digamos, 200 veces y sale cruz 80 veces, es razonable suponer que la probabilidad de que salga cruz al lanzar esa moneda se estima a 80/200, o sea 0.40. Queda claro que cuantas más veces se lance la moneda, más cercano a la probabilidad verdadera será el resultado que se obtenga.

Por eso, la definición de probabilidad utilizando este enfoque es:

$$P(X) = \lim_{K \to \infty} \frac{k}{K} \tag{3.2}$$

donde *K* es la cantidad de veces que se repite el experimento aleatorio y *k* es el número de veces en que ocurrió el evento *X*. Esta es la definición estricta: **la probabilidad de un evento es la frecuencia relativa que tendría en una serie infinita de realizaciones del experimento aleatorio.**

Postulados de la teoría de probabilidades

Los postulados básicos de la teoría de probabilidades son los siguientes.

I. La probabilidad de un evento A, $P(A)$, es un valor numérico que se encuentra en el intervalo [0,1]. Es decir,

$$0 \leq P(A) \leq 1.$$

II. La probabilidad de la totalidad del espacio muestral es igual a 1:

$$P(S) = 1.$$

III. Dados dos eventos mutuamente excluyentes M y N, pertenecientes al espacio muestral S, la probabilidad de la ocurrencia de uno u otro de ellos es igual a:

$$P(M \cup N) = P(M) + P(N).$$

IV. Si M y N son dos eventos no mutuamente excluyentes definidos en un mismo espacio muestral, entonces:

$$P(M \cup N) = P(M) + P(N) - P(M \cap N)$$

Ejemplo

Se extrae al azar una carta de una baraja de 52 naipes ¿Cuál es la probabilidad de que dicha carta sea una figura (F) o un corazón (C)?
$P(C) = (1/4)$
$P(F) = (12/52) = (3/13)$ y
$P(F \cap C) = (3/52)$ puesto que hay 3 figuras de corazones.
Finalmente:
$P(F \cup C) = P(F) + P(C) - P(F \cap C) = (3/13) + (1/4) - (3/52) = (11/26)$.

Este postulado puede ser aplicado a cualquier cantidad de eventos. Por ejemplo, para el caso de 3 eventos, A, B y C:
$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$.

Ejemplo

Un instituto de enseñanza de nivel medio ofrece cursos de 3 materias simultáneamente para 240 estudiantes: Matemática (M), Física (F) e Informática (I). Un total de 50 estudiantes cursan Matemática, 25 cursan Física, 18 cursan Informática, 12 cursan Matemática y Física, 10 cursan Matemática e Informática, 5 cursan Física e Informática y 3 cursan las 3 materias. ¿Cuál será la probabilidad de que un alumno elegido al azar curse, por lo menos, una de las tres materias?

$$\begin{aligned} P(A \cup B \cup C) &= \frac{50}{240} + \frac{25}{240} + \frac{18}{240} - \frac{12}{240} - \frac{10}{240} - \frac{5}{240} + \frac{3}{240} \\ &= \frac{69}{240} \\ &= 0.2875 \end{aligned}$$

V. Sea X^C el evento complementario del evento X, es decir que los eventos X^C y X son mutuamente excluyentes y colectivamente exhaustivos Entonces,

$$P(X^C) = 1 - P(X),$$

Ejemplo

Supongamos que se extrae una carta de una baraja inglesa. ¿Cuál es la probabilidad de que no sea un rey? Hay 4 reyes en la baraja así que la probabilidad de *rey* es igual a (4/52) = (1/13). Por tanto, aplicando el teorema vemos que la probabilidad de que la carta extraída no sea un rey será igual a 1 – (1/13) = (12/13)

Combinatoria

Repasaremos algunas operaciones básicas de conteo.

Permutaciones

Una permutación de un número de objetos es una disposición de estos objetos en un orden definido. El número de permutaciones de un conjunto de N elementos, tomados todos juntos es igual a $N!$ Designando este número por ${}_NP_N$, obtenemos que ${}_NP_N = N!$ donde $N!$ se lee "N factorial" y es el producto de todos los números enteros desde 1 hasta N, es decir: $N! = 1 \cdot 2 \cdot 3 \cdots (N-2) \cdot (N-1) \cdot N$. En particular, **1! = 1**; **2! = 1 2! = 2**, **3! = 1 2 3 = 6**, **4! = 1 2 3 4 = 24**. Finalmente, definimos $0! = 1$.

El número total de disposiciones de N objetos tomados de a n cada vez, con $n \leq N$, es:

$${}_NP_n = \frac{N!}{(N-n)!}.$$

Ejemplo. Cuatro banderas de señales han de ser izadas, una encima de la otra, en un mástil. ¿Cuántas señales diferentes pueden ser transmitidas izando 6 banderas diferentes de a 4 cada vez?:

$${}_6P_4 = \frac{6!}{(6-4)!} = \frac{2!\,3 \cdot 4 \cdot 5 \cdot 6}{2!} = 360.$$

Combinaciones

Una *combinación* es una selección de objetos considerados sin relación con su orden. El número total de combinaciones de un conjunto de N elementos tomados de a n cada vez, es:

${}_NC_n$ o $\begin{pmatrix} N \\ n \end{pmatrix}$ y es igual a: ${}_NC_n = \begin{pmatrix} N \\ n \end{pmatrix} = \dfrac{N!}{n!(N-n)!}.$

Por ejemplo, ¿de cuántas maneras distintas se pueden elegir 3 letras tomándolas de a 2 cada vez?:

$${}_3C_2 = \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \frac{{}_3P_2}{2!} = \frac{\frac{3!}{(3-2)!}}{2!} = \frac{3!}{2!(3-2)!} = 3$$

Es importante recordar que **en una permutación el orden cuenta** mientras que **en una combinación, el orden no cuenta**.

Ejemplo. Un equipo de básquet que está de viaje tiene 10 jugadores. El entrenador debe escoger un equipo inicial para el próximo juego. ¿Cuántos equipos diferentes de 5 jugadores pueden ser designados para este objetivo? Aquí no nos interesan las posiciones de cada uno de los 5 jugadores en cada equipo. Por tanto, es un problema de combinaciones, y:

$${}_{10}C_5 = \frac{10!}{5!(10-5)!} = 252$$

Si al escoger un equipo, el entrenador también designa las posiciones, entonces el orden cuenta y el problema es de permutaciones:

$${}_{10}P_5 = \frac{10!}{(10-5)!} = 30240$$

Probabilidades condicionales

Cuando se reúne información adicional a la que se disponía inicialmente, el espacio muestral puede resultar redimensionado. Es decir, cuando hay una reducción de la incertidumbre (ya sea por aumento en la información disponible o por el empleo de supuestos por parte del ingeniero), puede que haya puntos muestrales que desaparezcan del espacio muestral resultando éste, reducido. Por ejemplo, frente al experimento aleatorio de lanzar un dado homogéneo, decimos que la probabilidad de que salga un 5 es igual a 1/6. Ahora, si se ha lanzado el dado pero solamente se puede saber que ha salido un número impar, pero no qué número ha salido, sigue habiendo una grado de incertidumbre, pero no cabe duda de que dicha incertidumbre es menor puesto que ya se sabe que salió un número impar: el nuevo espacio muestral es, ahora, $S = \{1,3,5\}$. Ahora, la probabilidad de que el dado haya salido 5 ya no es 1/6 sino 1/3. Las probabilidades calculadas en espacios muestrales reducidos por información o supuestos adicionales se denominan **probabilidades condicionales.** Veamos un ejemplo.

Supongamos que se toma una muestra de 100 estudiantes y a cada uno de ellos se le hacen dos preguntas: (1) si ha aprobado el curso de Estadística y, (2) si le gustan las carreras de autos. Los resultados de la encuesta son los siguientes:

Cuadro 3.1.

	Le gustan las carreras de autos	No le gustan las carreras de autos	Total
Aprobó Estadística	28	52	80
No aprobó Estadística	12	8	20
Total	40	60	100

Se elige un estudiante al azar en dicha muestra y se definen dos eventos: X (el estudiante aprobó Estadística) e Y (al estudiante le gustan las carreras de autos). Entonces:

$$P(X) = \frac{80}{100} = 0.80 \text{ y } P(Y) = \frac{40}{100} = 0.40 \cdot$$

Ahora, supongamos que la elección es realizada entre los estudiantes que han aprobado Estadística. Entonces, ¿cuál es la probabilidad de que el estudiante elegido sea afecto a las carreras de autos? Hay una información adicional que cambia el espacio muestral: se está dando por cierto que el estudiante aprobó Estadística y la única incertidumbre que queda es si le gustan las carreras de autos o no le gustan. Por tanto, el nuevo espacio muestral está restringido a la segunda fila del cuadro: S = {le gustan las carreras, no le gustan las carreras} con un tamaño igual a 80 (el total de estudiantes que aprobaron Estadística). Entonces, la probabilidad buscada es:

$$P(Y/X) = \frac{28}{80} = 0.35 \cdot$$

De la misma manera se pueden calcular otras probabilidades condicionales como, por ejemplo, la probabilidad de que haya aprobado Estadística dado que le gustan las carreras de autos. En ese caso:

$$P(X/Y) = \frac{28}{40} = 0.70 \cdot$$

También se podría haber transformado todo el cuadro en probabilidades, dividiendo por el total:

Cuadro 3.2.

	Le gustan las careras de autos	No le gustan las careras de autos	Total
Aprobó Estadística	(28/100) = 0.28	(52/100) = 0.52	**(80/100) = 0.80**
No aprobó Estadística	(12/100) = 0.12	(8/100) = 0.08	**(20/100) = 0.20**
Total	**(40/100) = 0.40**	**(60/100) = 0.60**	**(100/100) = 1.00**

y calcular las probabilidades condicionales de la siguiente manera:

$$P(Y/X) = \frac{0.28}{0.80} = 0.35 \text{ y}$$

$$P(X/Y) = \frac{0.28}{0.40} = 0.70 \cdot$$

Las probabilidades que están en los márgenes del cuadro (0.80, 0.20, 0.40 y 0.60) se denominan **probabilidades marginales** y las probabilidades que están en el cuerpo del cuadro (0.28, 0.52, 0.12 y 0.08) se denominan **probabilidades conjuntas.**

En general, se presentan tres tipos de problemas:

(i) se conoce la probabilidad conjunta de dos eventos y una de las probabilidades marginales y se desea conocer una probabilidad condicional; es el caso del ejemplo de más arriba, donde se conocen las probabilidades conjuntas y las marginales y, con eso, se pueden calcular probabilidades condicionales;

(ii) se conoce una probabilidad condicional y una probabilidad marginal y se desea calcular una probabilidad conjunta;

(iii) se conoce una probabilidad condicional y una probabilidad conjunta y se desea calcular una probabilidad marginal.

La ecuación correspondiente al caso (i) es:

$$P(Y/X) = \frac{P(X \cap Y)}{P(X)} \text{ ó}$$

$$P(X/Y) = \frac{P(X \cap Y)}{P(Y)} \qquad (3.3)$$

Para el caso (ii):

$$P(X \cap Y) = P(Y/X)\ P(X) \text{ ó}$$

$$P(X \cap Y) = P(X/Y)\ P(Y) \qquad (3.4)$$

Y para el caso (iii):

$$P(X) = \frac{P(X \cap Y)}{P(Y/X)} \quad \text{y} \quad P(Y) = \frac{P(X \cap Y)}{P(X/Y)} \qquad (3.5)$$

Eventos independientes

Se dice que dos eventos son **estadísticamente independientes** cuando la ocurrencia de uno de ellos no afecta la probabilidad de ocurrencia del otro y, entonces, la probabilidad de su ocurrencia simultánea (probabilidad conjunta) es igual al producto de sus probabilidades individuales: $P(X \cap Y) = P(X)\ P(Y)$. Contrariamente, si la ocurrencia de uno de los eventos afecta la probabilidad de la ocurrencia del otro, entonces se dice que esos dos eventos son **estadísticamente dependientes** y, entonces, su probabilidad conjunta es igual al producto de la ocurrencia de uno de ellos por la probabilidad condicional de la ocurrencia del segundo dado que ha ocurrido el primero: $P(X \cap Y) = P(X)\ P(Y/X)$.

Ejercicios

3.1. Un turno de exámenes consta de 5 fechas diferentes. Un alumno debe rendir 3 materias. ¿De cuántas maneras diferentes se puede anotar para rendir sus exámenes si sólo puede rendir una materia por fecha? 60

3.2. En un estudio sobre la regeneración de la palmera yatay (Butia yatay) en el Parque Nacional El Palmar se registró la supervivencia de 200 plántulas de palmera tomadas al azar dentro de un palmar. Entre las plántulas elegidas, 120 estaban ubicadas a menos de 4 metros de distancia de la palmera adulta más cercana (bajo su copa) y 80 estaban ubicadas a más de 4 m de distancia de la palmera adulta más cercana. Al cabo de un año, seguían vivas 80 de las plántulas ubicadas a menos de 4 m de una palmera adulta y 60 de las ubicadas a más de 4 m de la palmera adulta más cercana. Definamos ahora el experimento aleatorio que consiste en tomar al azar una de las plántulas:

a. *¿Qué eventos simples componen el espacio muestral de este experimento?*
b. *Señalar dos eventos mutuamente excluyentes en dicho espacio. ¿Cuál es la probabilidad de cada uno? ¿Cual es la probabilidad de que ocurra uno o el otro?*
c. *Señalar dos eventos NO mutuamente excluyentes. ¿Cuál es la probabilidad de cada uno?*
d. *¿Cuál es la probabilidad de que una plántula tomada al azar haya sobrevivido?* 0,7
e. *¿Cuál es la probabilidad de que una plántula ubicada a más de 4 m de distancia de la palmera adulta más cercana haya muerto?* 0,75
f. *¿Es independiente la supervivencia de las plántulas estudiadas de su ubicación respecto de la palmera adulta más cercana? ¿Por qué?*

3.3. En una planta procesadora de frutas dos inspectores revisan visualmente la fruta. Cuando aparece una fruta defectuosa, la probabilidad de que no sea detectada por el primer inspector es igual a 0.1. De aquellas no detectadas por el primer inspector, el segundo inspector sólo detecta 5 de cada 10.

a. ¿Cuánto vale la probabilidad de que una fruta defectuosa no son detectadas por ninguno de los inspectores? 0,05
b. Explicar esta probabilidad en términos de la definición estricta de probabilidad. $P[A] = 0,1$ $P[B \cap A] = 0.05$

3.7 A continuación se muestra una tabla probabilística acerca del nivel de instrucción de productores de una zona y la implementación de nuevas técnicas de cultivo y sea *A* el evento *nivel de instrucción bajo* y B, el evento *no implementa nuevas técnicas de cultivo.*

		Implementación de nuevas técnicas de cultivo	
		No	Sí
Nivel de instrucción	Bajo	0.40	0.20
	Alto	0.10	0.30

a. *Calcular P(A∪B).*
b. *¿Son independientes el nivel de instrucción de los productores de esa zona y la implementación de nuevas técnicas de cultivo?*

3.5 La siguiente tabla muestra algunas de las probabilidades relacionadas con el aumento de peso de pollos criados con distintas raciones. Además, se sabe que el 80% de los pollos criados con la ración A aumentan menos de 50 g /día,:

Aumento de peso	Tipo de Ración			
	A	B	C	
> 50 gr. / día			0.6	0.7
< 50 gr. / día				
	0.10	0.20		

a. *Completar la Tabla.*
b. *¿Cuál es la probabilidad de encontrar pollos con aumentos de peso menores a 50 g /día y alimentados con la ración B?*
c. *¿Cuál es la probabilidad de que pollos alimentados con la ración B aumenten menos de 50 g /día?*
d. *¿Cuál es la probabilidad de que pollos alimentados con la ración C aumenten menos de 50 g /día?*
e. *¿El aumento de peso es independiente del tipo de ración que reciben los pollos? Justificar la respuesta.*

3.6. En un estudio sobre el control de la fusariosis del trigo (una enfermedad producida por un hongo), se pusieron a prueba tres tipos de dispositivo aspersor (boquillas) para aplicar un fungicida (A, B y C). Para ello se seleccionaron 80 cultivos, cada uno fue tratado con un tipo de boquilla asignado al azar y un tiempo después se registró la presencia o ausencia de la enfermedad en cada uno. Entre los 80 cultivos tratados, sólo 15 presentaron la enfermedad. El número de cultivos infectados tratados con las boquillas A y C fue igual y equivalente a un tercio del número de cultivos infectados tratados con la boquilla B. Además, entre los cultivos tratados con la boquilla A, la mitad apareció infectada.

a. *¿Cuál fue la probabilidad de contraer fusariosis de los cultivos tratados con la boquilla A?*
b. *¿Cuál fue la probabilidad de no contraer fusariosis de los cultivos tratados las boquillas B o C?*

3.7 En un lago conviven dos especies de pejerrey (A y B) en igual proporción. El 22% de los pejerreyes de la especie A y el 35% de los de la especie B están infectados por un protozoo intestinal.

a. *¿Cuál es la probabilidad de que un pejerrey tomado al azar esté infectado?*

b. *¿Cuál es la probabilidad de que un pejerrey tomado al azar esté infectado y además pertenezca a la especie A?*

c. *¿Cuál es la probabilidad de que un pejerrey infectado pertenezca a la especie A?*

d. *Usando probabilidades condicionales, explicar por qué la infección con el protozoo intestinal no es estadísticamente independiente de la especie de pejerrey.*

P [A/I] = P [B/I]

0,38 ≠ 0,7

Salud \ Esp.	A	B	
INF	0,11	0,175	0,285
No INF	0,39	0,325	0,715
	0,5	0,5	1

P. Marginal

3.5 c P [B ∩ >50] = P [>50] . P [B]

DISTRIBUCIONES DE PROBABILIDADES

Variables aleatorias

En la aplicación de las probabilidades para el análisis de la información proveniente de experimentos aleatorios, se trabaja con variables definidas a partir de los espacios muestrales. Dichas variables reciben el nombre de **variables aleatorias**. Las variables aleatorias, dado que provienen de un espacio muestral, son variables que pueden asumir un determinado conjunto de valores diferentes con determinadas probabilidades. Los análisis estadísticos involucran a la **distribución de probabilidades** de la variable aleatoria de interés. Estas variables aleatorias pueden ser de dos clases: variables aleatorias **discretas** y variables aleatorias **continuas**.

Variables aleatorias discretas

Las variables aleatorias discretas sólo pueden tomar valores nominales o valores cuantitativos discretos. A cada uno de esos valores o categorías, le corresponderá una probabilidad. Así queda constituida la distribución de probabilidades de la variable aleatoria discreta. La suma de las probabilidades correspondientes a todos los valores o categorías que puede tomar de una variable aleatoria discreta es igual a 1.

Ejemplos:

- Sea la variable aleatoria *estado sanitario de un animal* con dos categorías, *sano* (*H*) y *enfermo* (*E*). Entonces, el espacio muestral es $S = \{H,E\}$.

- Sea la variable aleatoria *número de puntos obtenidos al arrojar un dado equilibrado*. Entonces, el espacio muestral es $S = \{1, 2, 3, 4, 5, 6\}$. La distribución de probabilidades correspondiente a esta variable es

Nº de puntos	1	2	3	4	5	6
Probabilidad	1/6	1/6	1/6	1/6	1/6	1/6

Entonces, escribimos: $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = (1/6)$. Si denotamos a una variable aleatoria discreta con el símbolo X, y a cada uno de los valores particulares que puede tomar x_i, entonces, la probabilidad de un valor particular x_i, se denotará $P(x_i)$.

Las distribuciones de probabilidades de variables discretas se pueden representar gráficamente mediante un diagrama de barras verticales en el cual se inscriben los distintos valores que la variable aleatoria puede tomar en el eje de abscisas y sus respectivas probabilidades en el eje de ordenadas.

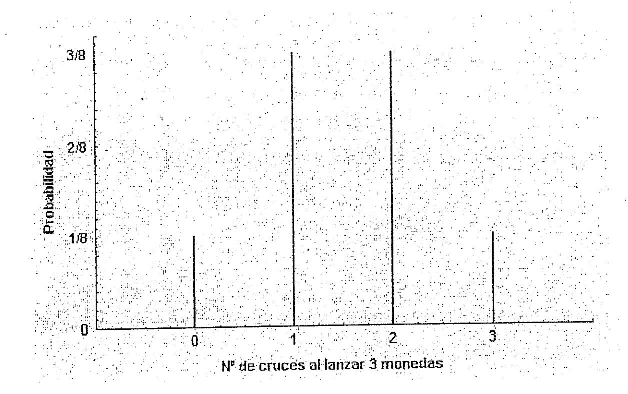
Ejemplo:

- Sea la variable aleatoria discreta *número de cruces que se pueden obtener al lanzar tres monedas equilibradas*. Si los lanzamientos de las tres monedas son eventos independientes, la distribución de probabilidades de esta variable aleatoria es la siguiente:

Nº de cruces en 3 monedas	0	1	2	3
Probabilidad	1/8	3/8	3/8	1/8

La representación gráfica de esta distribución de probabilidades se muestra en la Figura 4.1.:

Figura 4.1. Distribución de variable aleatoria discreta.



Distribución de probabilidades acumulativa

Muchas veces es necesario conocer la probabilidad, no ya de un suceso puntual particular, sino de un conjunto de sucesos y, entonces, surge la necesidad de acumular probabilidades. Aquí estamos tratando otra vez con ese concepto y por eso presentamos la **distribución de probabilidades acumulativa** de una variable aleatoria discreta como la probabilidad de que la variable aleatoria asuma un valor tope o menor, es decir, interesa conocer la probabilidad $P(X \leq x_i)$.

Ejemplo.
Utilizando otra vez el ejemplo anterior, se desea conocer la probabilidad de que al lanzar 3 monedas equilibradas, se obtengan, *a lo sumo*, 1 cruz. Entonces:

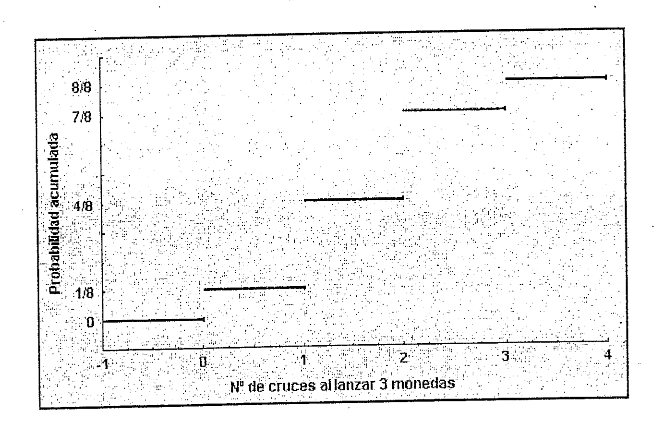
$$\begin{aligned} P(X \leq 1) &= P(X=0) + P(X=1) \\ &= (1/8) + (3/8) \\ &= (4/8) \\ &= (1/2). \end{aligned}$$

En el siguiente cuadro se representan la distribución de probabilidades y la distribución de probabilidades acumulativa del experimento de lanzar 3 monedas:

Nº de cruces en 3 monedas	0	1	2	3
Probabilidad	1/8	3/8	3/8	1/8
Probabilidad acumulada	1/8	4/8	7/8	8/8

Y la representación gráfica de la distribución de probabilidades acumuladas es:

Figura 4.2. Distribución de probabilidades acumuladas.



Esperanza matemática o media poblacional de una variable aleatoria discreta

La esperanza matemática de una variable aleatoria discreta es el promedio de todos los valores que tomaría en una serie infinita de experimentos aleatorios. Como por definición, la frecuencia relativa de cada valor de la variable infinita es justamente su probabilidad la esperanza matemática puede ser calculada como:

$$E(X) = \sum_i x_i \cdot P(x_i) \tag{4.1}$$

La esperanza matemática de X se suele simbolizar μ_X y es también llamada valor esperado de X o media poblacional de X.

Ejemplos:

- Para el caso del lanzamiento de las 3 monedas:

$$\begin{aligned} E(X) &= \mu_X \\ &= 0\ (1/8) + 1\ (3/8) + 2\ (3/8) + 3\ (1/8) \\ &= (3/2). \end{aligned}$$

- Para el caso de la variable aleatoria *número de puntos obtenidos al arrojar un dado equilibrado*:

$$\begin{aligned} E(X) &= \mu_X \\ &= 1\ (1/6) + 2\ (1/6) + 3\ (1/6) + 4\ (1/6) + 5\ (1/6) + 6\ (1/6) \\ &= 3.5. \end{aligned}$$

Propiedades de esperanza matemática de una variable discreta:

I. Sea k una constante arbitraria. Entonces, si se suma k a cada uno de los valores de una variable aleatoria X, resulta:

$$E(X + k) = E(X) + k. \tag{4.2}$$

Ejemplo.
Si sumamos la constante 2 a la variable aleatoria *número de puntos obtenidos al arrojar un dado equilibrado*, resulta: $E(X + 2) = E(X) + 2$. En efecto:

$$\begin{aligned} E(X + 2) &= \mu_{X+2} \\ &= (1+2)\ (1/6) + (2+2)\ (1/6) + (3+2)\ (1/6) + (4+2)\ (1/6) + \\ &\quad + (5+2)\ (1/6) + (6+2)\ (1/6) \\ &= 3\ (1/6) + 4\ (1/6) + 5\ (1/6) + 6\ (1/6) + 7\ (1/6) + 8\ (1/6) \\ &= (33/6) \\ &= 5.5 \\ &= \mathbf{3.5 + 2.} \end{aligned}$$

II. Sea k una constante arbitraria. Entonces, si multiplica por k a cada uno de los valores de una variable aleatoria X, resulta:

$$E(X\ k) = E(X) \cdot k. \tag{4.3}$$

Ejemplo.
Si multiplicamos por la constante 2 a la variable aleatoria *número de puntos obtenidos al arrojar un dado equilibrado*, resulta: $E(X\ 2) = E(X)\ 2$. En efecto:

$$
\begin{aligned}
E(X \cdot 2) &= \mu_{X \cdot 2} \\
&= (1 \cdot 2)\ (1/6) + (2 \cdot 2)\ (1/6) + (3 \cdot 2)\ (1/6) + (4 \cdot 2)\ (1/6) + \\
&\quad + (5 \cdot 2)\ (1/6) + (6 \cdot 2)\ (1/6) \\
&= 2\ (1/6) + 4\ (1/6) + 6\ (1/6) + 8\ (1/6) + 10 \cdot (1/6) + 12\ (1/6) \\
&= (42/6) \\
&= 7.0 \\
&= (3.5) \cdot 2.
\end{aligned}
$$

III. Juntando las dos ecuaciones 4.2. y 4.3. en una sola, obtenemos que :

$$E(k_1\ X + k_2) = k_2 + k_1\ E(X) \tag{4.4}$$

donde k_1 y k_2 son constantes arbitrarias. Se deja como ejercicio para el lector, aplicar esta última propiedad a la variable aleatoria *número de puntos obtenidos al arrojar un dado equilibrado*.

IV. La ecuación 4.1. implica que si la esperanza de una variable aleatoria X es $E(X) = \mu$; entonces

$$E(X - \mu) = 0 \tag{4.5}$$

Variancia poblacional de una variable aleatoria discreta

Similarmente a lo apuntado en el capítulo de Descripción de la Información, la variancia de una variable aleatoria mide la dispersión de los valores que toma en la población alrededor de su esperanza matemática. La variancia de una variable aleatoria discreta X se define como:

$$
\begin{aligned}
V(X) &= \sigma_X^2 \\
&= \sum_i \left[p(x_i) \cdot (x_i - \mu)^2 \right]
\end{aligned}
\tag{4.6}
$$

Comparando la fórmula 4.2. con la 4.6. es posible visualizar que la variancia poblacional no es otra cosa que $V(X) = E[(X-\mu)^2]$, el valor esperado de los cuadrados de los desvíos de X con respecto a su media poblacional μ.

Propiedades de la variancia de una variable aleatoria discreta:

I. Si sumamos una constante a una variable aleatoria, su variancia no altera.

$$
\begin{aligned}
V(X + k) &= \\
&= E\ [(X + k - E(X+k))^2] \\
&= E\ [(X + k - E(X) - k)^2] \\
&= E\ [(X - E(X)\ )^2] \\
&= V(X)
\end{aligned}
\tag{4.7}
$$

II. Si multiplicamos una variable aleatoria por una constante, su variancia resulta multiplicada por dicha constante elevada al cuadrado. En efecto

$$
\begin{aligned}
V(k\ X) &= \\
&= E\ [(k\ X - E(k\ X))^2] \\
&= E\ [(k\ X - k\ E(X))^2] \\
&= E\ [k^2\ (X - E(X))^2] \\
&= k^2\ E\ [(X - E(X))^2] \\
&= k^2\ V(X)
\end{aligned}
\tag{4.8}
$$

III. Combinando las propiedades (i) y (ii) resulta que

$$V(k_1\ X + k_2) = k_1^2\ V(X) \tag{4.9}$$

Desvío standard poblacional y coeficiente de variación

El **desvío standard** poblacional es simplemente la raíz cuadrada de la variancia poblacional y el **coeficiente de variación** es el cociente del desvío standard sobre la esperanza matemática, multiplicado por 100.

Ejemplo.
Siguiendo con la variable aleatoria discreta *número de puntos obtenidos al arrojar un dado equilibrado*, resulta:

$$\begin{aligned}\sigma &= \sqrt{V(X)} \\ &= \sqrt{\frac{35}{12}} \\ &= 1.708, \qquad y \end{aligned} \tag{4.10}$$

$$\begin{aligned} cv &\approx \frac{1.708}{3.5} \cdot 100 \\ &= 48.8 \end{aligned}$$

Variables aleatorias continuas

Las variables aleatorias continuas toman valores en el campo de los números reales y, por lo tanto, su distribución de probabilidades está representada por una función continua puesto que la variable puede tomar infinitos valores.

Ahora, dada esta característica de continuidad, la probabilidad de que la variable X tome un valor particular infinitesimalmente exacto, es igual a 0. Esto nos obliga a que, cuando se trata de variables aleatorias continuas, tengamos que calcular probabilidades de intervalos entre dos valores y no para un dado valor único. Ya no escribiremos $P(X = x_i)$ sino $P(X \leq x_i)$ o $P(x_i \leq X \leq x_j)$. Así que, ahora, la probabilidad resultará ser un área en la representación gráfica y estará determinada por una integral bajo la curva de una función que se denomina **función de densidad de probabilidad**, $f(x)$. En el siguiente gráfico se esquematizan estos conceptos:

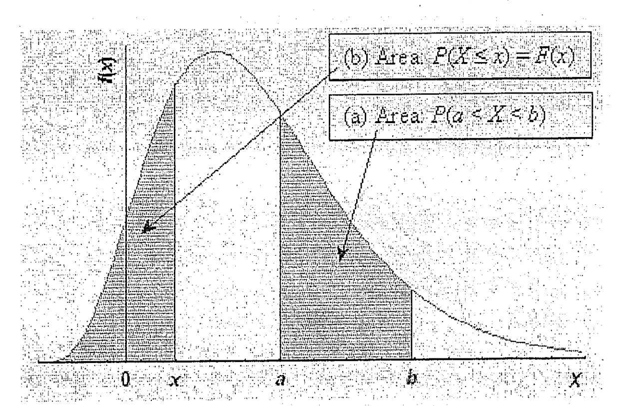


Figura 4.3. Curva de la función de densidad de probabilidad.

De manera que cuando calculamos probabilidades para variables aleatorias continuas, estamos calculando probabilidades **acumuladas**. Simbolizaremos las probabilidades de que la variable X sea menor o igual a un valor particular x como $F(x)$ que es, como se dijo antes, la integral de la función de densidad $f(x)$, desde $-\infty$ hasta x, es decir, $F(x) = P(X \leq x)$. $F(x)$ se denomina **función de distribución de probabilidades**. Asimismo, para un intervalo $[x_1,x_2]$, resulta que $P(x_1 < X < x_2) = F(x_2) - F(x_1)$. Todas estas consideraciones nos llevan a la conclusión de que el área total bajo la curva de la función de densidad (que representa, en este caso, la probabilidad de todo el espacio muestral) debe ser, necesariamente, igual a 1.

Para ser función de densidad, una función debe cumplir dos requisitos fundamentales:

(a) debe ser no negativa en todo su intervalo de definición;
(b) la integral definida de la función calculada sobre todo el intervalo de definición debe ser igual a 1.

Para ilustrar las propiedades de las distribuciones de probabilidad de variables continuas presentamos a continuación un ejemplo en el cual la función de densidad de probabilidad es una función lineal.

Ejemplo.
Sea la variable aleatoria *X* cuya función de densidad se define de la siguiente manera: en el intervalo [2;4], *f*(*x*) = (1/18) (3 + 2*x*), y para cualquier otro valor de *x*, *f*(*x*) = 0.

- Primeramente, observemos que *f*(*x*) es continua en [2;4].
- En segundo término, podemos ver que *f*(*x*) > 0 en [2;4].
- Finalmente, veamos que el área total bajo la curva de la función de densidad es, efectivamente, igual a 1:

$$\int_{-\infty}^{+\infty} f(x)dx$$

$$= \int_{-\infty}^{2} 0 \cdot dx + \int_{2}^{4} \frac{1}{18} \cdot (3+2x) \cdot dx + \int_{4}^{+\infty} 0 \cdot dx$$

$$= \frac{1}{18} \cdot \int_{2}^{4} (3+2x)dx$$

$$= \frac{1}{18} \cdot \left(3x + x^2\right)\Big|_2^4$$

$$= \frac{1}{18} \cdot \left[(12+16) - (6+4)\right] = 1$$

Gráficamente:

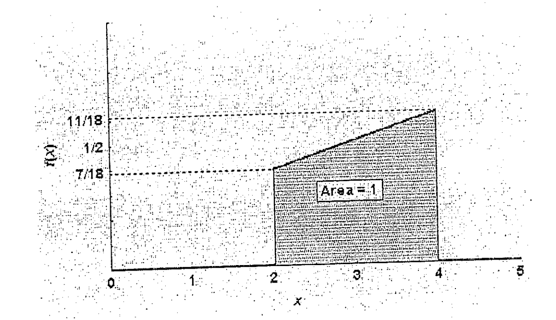


Figura 4.4. El área que queda debajo de la función de densidad de probabilidad en todo el espacio muestral es igual a 1

Ahora determinaremos la función de distribución *F*(*x*):

$$\int_{2}^{x} \frac{1}{18} \cdot (3+2t)dt$$

$$= \frac{1}{18} \cdot \left(\left[3 \cdot t\right]_2^x + \left[t^2\right]_2^x\right)$$

$$= \frac{1}{6} \cdot (x-2) + \frac{1}{18} \cdot \left(x^2 - 4\right)$$

$$= \frac{x^2}{18} + \frac{x}{6} - \frac{5}{9}$$

Finalmente calculemos una probabilidad particular, por ejemplo, $P(2.5 \leq X \leq 3.5)$.

$$P(2.5 \leq X \leq 3.5) = F(3.5) - F(2.5)$$
$$= 0.708333 - 0.208333$$
$$= 0.5.$$

Gráficamente:

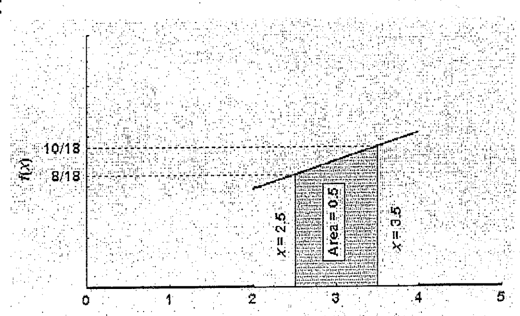


Figura 4.5.

Esperanza y variancia poblacionales de una variable aleatoria continua

Sea la variable aleatoria X con función de densidad $f(x)$ definida en el intervalo $[a,b]$. Entonces su **esperanza matemática o media** se define como:

$$E(X) = \mu = \int_a^b x \cdot f(x) \cdot dx \qquad (4.11)$$

y su **variancia**, como:

$$V(X) = \sigma^2 = \int_a^b (x-\mu)^2 f(x) \cdot dx \qquad (4.12)$$

Las propiedades de la esperanza (4.2., 4.3, 4.4 y 4.5) y de la variancia (4.7, 4.8 y 4.9) de una variable aleatoria discreta se aplican para variables aleatorias continuas, es decir:

I $E(X + k) = E(X) + k$;
II $E(X\,k) = E(X) \cdot k$;
III $E(k_1 X + k_2) = k_2 + k_1 E(X)$, donde k_1 y k_2 son constantes arbitrarias;
IV Sea la variable aleatoria X con $E(X) = \mu$; entonces $E(X - \mu) = 0$.

La desigualdad de Tchebysheff

Una importante propiedad de las distribuciones de probabilidad es la que se conoce como desigualdad de Tchebysheff. Según esta propiedad, dada una variable aleatoria x, con esperanza $E(x) = \mu$ y variancia σ^2, se cumple que:

$$P(|x-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

o, puesto de otra manera,

$$P(|x-\mu| < k\sigma) \geq 1-\frac{1}{k^2}, \qquad k \geq 1$$

Esto significa que dada cualquier variable aleatoria con esperanza μ y variancia σ^2 *sin importar cuál sea la forma de su distribución*, la probabilidad de que esta se desvíe de su media poblacional (o esperanza) para un lado o para otro en una cantidad mayor a k veces su desvío standard es limitada; no puede pasar de $1/k^2$. Otra manera de escribir esta propiedad es:

$$P(\mu - k\sigma < x < \mu + k\sigma) > 1-\frac{1}{k^2}$$

Esta última expresión indica cómo puede utilizarse σ como medida de dispersión y puede ser usada en una gran variedad de casos para lo cual sólo basta con el supuesto de que μ y σ^2 existen y son finitas, es decir, no se hace ninguna suposición acerca de la forma de la distribución de la variable aleatoria en la población. La desigualdad de Tchebysheff permite *acotar* la probabilidad de que las probabilidades de los valores de la variable aleatoria queden a una dada distancia de su media. Sin embargo, no debe olvidarse de que se trata sólo de una *cota*, es decir, un valor máximo de probabilidad.

Ejemplo.
El tiempo que tarda un camión en ser cargado completamente, sigue una distribución que tiene media μ = 50 minutos y variancia σ^2 = 100 minutos^2. Utilizaremos la desigualdad de Tchebysheff para calcular la probabilidad de que el tiempo de carga esté entre 30 y 70 minutos.

Según la desigualdad de Tchebysheff : $P(|x-\mu| \geq k\sigma) \leq \frac{1}{k^2}$;

haciendo $t = k\sigma$, esto implica que: $P(|x-\mu| \geq t) \leq \frac{\sigma^2}{t^2}$.

Ahora, el problema requiere que $30 \leq t \leq 70$ lo que implica, sabiendo que $\mu = 50$, que:

$P(30 \leq t \leq 70) = 1 - P(|x-50| \geq 20)$. Aplicando el teorema obtenemos que:

$$P(|x-\mu| \geq t) \leq \frac{\sigma^2}{t^2} = \frac{10^2}{20^2} = 0.25 \Rightarrow P(30 \leq t \leq 70) \geq 1-0.25 = 0.75$$

En términos del problema diremos que la probabilidad de que el tiempo de carga del camión esté entre 30 y 70 minutos es, por lo menos, de 0.75.

Variables aleatorias estandarizadas

Como veremos más adelante, muchas veces resulta conveniente trabajar con las variables transformadas a través del proceso de **estandarización**, en lugar de hacerlo con las variables originales. El proceso de estandarización consiste, simplemente, en transformar cada uno de los valores de la variable restándole la

media aritmética (o sea, la esperanza matemática) y dividiendo dicha resta por el desvío standard. La nueva variable se simboliza con la letra Z:

$$Z = \frac{X - \mu_X}{\sigma_X} \qquad (4.13)$$

La nueva variable se denomina **variable aleatoria estandarizada** o **variable aleatoria standard** y, dado que surge de restar la media y dividir por el desvío standard, tendrá media igual a 0 y variancia igual a 1, pues:

$$\begin{aligned} E(Z) &= E\left(\frac{X - \mu_X}{\sigma_X}\right) \\ &= \frac{E(X) - E(X)}{\sigma_X} \text{ y} \\ &= 0 \end{aligned}$$

La variancia de una variable estandarizada es siempre igual a 1 porque:

$$\begin{aligned} V(Z) &= V\left(\frac{X - \mu_X}{\sigma_X}\right) \\ &= \frac{1}{\sigma_X^2} \cdot V(X - \mu_X) \\ &= \frac{V(X)}{\sigma_X^2} \\ &= 1 \end{aligned}$$

Algunas distribuciones de probabilidades de uso común

Hay una gran cantidad de fenómenos naturales y sociales que se caracterizan por compartir un patrón de comportamiento similar. Además, se han descubierto modelos matemáticos sencillos que tienen la capacidad de describir muy ajustadamente dichos comportamientos. Por estas razones es que se ha consagrado su uso como herramienta de análisis. En este curso veremos un modelo para variables aleatorias discretas y tres modelos para variables aleatorias continuas, aunque poniendo especial énfasis en uno de ellos.

Un modelo de variable aleatoria discreta

La distribución binomial

Este modelo se emplea con variables aleatorias discretas que sólo pueden asumir dos valores o categorías que pueden denominarse de varias formas equivalentes: 0 y 1, *éxito* y *fracaso*, *defectuoso* y *no defectuoso*, etc., dependiendo del problema de que se trate. Uno de esos dos estados tiene una probabilidad constante que designaremos con la letra π y, por tanto, el otro estado alternativo tendrá una probabilidad 1 - π ya que es el evento complementario.

Este tipo de variables se denominan **dicotómicas** y su distribución de probabilidades se genera mediante la repetición de n experimentos aleatorios independientes, en cada uno de los cuales se mantienen constantes las probabilidades de los dos estados posibles de la variable aleatoria.

Ejemplos:

- La variable aleatoria *faz de una moneda* es un típico ejemplo de variable aleatoria dicotómica puesto que sólo puede asumir dos estados, *cara* y *cruz* y, por tanto, su distribución de probabilidades es bien descripta por la **distribución binomial**. Un proceso binomial con una moneda podría consistir en lanzar la moneda 20 veces y estudiar el número de cruces que han salido en esos 20 lanzamientos.
  La función de distribución de probabilidades binomial permite calcular la cantidad x de veces que se produce un dado resultado de una variable binomial, en n experimentos aleatorios independientes y se define de la siguiente manera:

$$b(x;n,\pi)=\binom{n}{x}\pi^{x}(1-\pi)^{n-x} \qquad (4.14)$$

  donde $\binom{n}{x}$ es un número combinatorio como hemos visto en la clase anterior, y π es la probabilidad (constante de experimento en experimento) del resultado buscado. De modo que una distribución de probabilidades binomial queda completamente definida conociendo los valores de n y π.

- En un lote de 12 plantas, 3 tienen flores púrpura. Si se extrae del lote una muestra al azar de 3 plantas, con reposición, ¿cuál es la probabilidad de que: (a) exactamente 1 planta tenga flores púrpura, y (b) a lo sumo 1 planta tenga flores púrpura? El muestreo con reposición asegura la independencia de las elecciones sucesivas, así que se puede aplicar el modelo binomial. Como π = (3/12) = 0.25, entonces:

1)
$$b(1;3,0.25)=\binom{3}{1}\cdot 0.25^{1}0.75^{3-1} = 0.42$$

y

2)
$$b(0;3,0.25)+b(1;3,0.25)=\binom{3}{0}\cdot 0.25^{0}0.75^{3}+\binom{3}{1}\cdot 0.25^{1}0.75^{2} = 0.84$$

Como toda distribución de probabilidades, la distribución binomial también permite calcular probabilidades acumuladas. La distribución de probabilidades acumuladas permite calcular la probabilidad de obtener a lo sumo m resultados en n ensayos:

$$\begin{aligned} B(m;n,\pi) &= P(X\le m) \\ &= b(0;n,\pi)+b(1;n,\pi)+\cdots+b(m;n,\pi) \\ &= \sum_{x=0}^{m} b(x;n,\pi) \end{aligned}$$

Ejemplo:

4.1. Una moneda equilibrada es arrojada 10 veces: ¿cuál es la probabilidad de obtener 8 o más caras (es decir, por lo menos 8 caras)? Aquí tenemos un modelo binomial con n = 10, π = 0.5. La probabilidad buscada es la de obtener 8, 9 ó 10 caras. Entonces:

$$\sum_{x=8}^{10} b(x;10,0.5) = 1 - \sum_{x=0}^{7} b(x;10,0.5)$$
$$\cong 1 - 0.94531$$
$$= 0.05469$$

Se puede demostrar que la esperanza matemática de una distribución binomial es igual a $n\ \pi$ y que su variancia es igual a $n\ \pi\ (1 - \pi)$. Por ejemplo, una moneda es lanzada 10 veces la esperanza del número de caras obtenidas es $E(X) = 10\ 0.5 = 5$ y la variancia es $V(X) = 10\ 0.5\ 0.5 = 2.5$.

Modelos de variables aleatorias continuas

La distribución normal

La distribución normal es el modelo de distribución de probabilidades más importante en aplicaciones relacionadas con la agronomía y las ciencias ambientales. Si una variable aleatoria X tiene una distribución normal, su función de densidad de probabilidad es:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

La curva descripta por esta función es la que se ve en la Figura 4.6 Los **parámetros** que definen la distribución de probabilidad de esta variable X son su **media** (μ) y su **variancia** (σ^2). Conociendo la media y la variancia de una variable aleatoria que tiene una distribución normal se conoce completamente su distribución. Una forma especial de la distribución normal es la **distribución normal standard** que resulta de restar, a cada uno de los valores de la variable, la media y el desvío standard de la distribución, como hemos visto algunas páginas atrás:

$$Z = \frac{X - \mu_X}{\sigma_X} \quad (4.15)$$

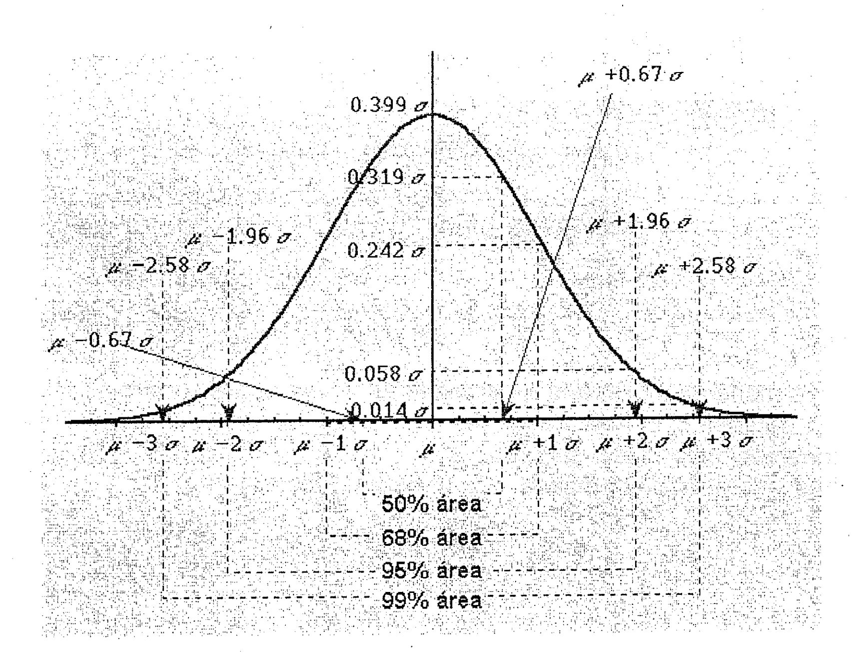


Figura 4.6. Distribución Normal, curva de densidad de probabilidad. Las áreas que quedan bajo la curva miden probabilidades de que una variable con distribución normal tome valores en los intervalos correspondientes.

Como ocurre con toda distribución de probabilidades, el área bajo la curva de la función de densidad, es igual a 1 (es la probabilidad de la totalidad del espacio muestral). Además, la función es perfectamente simétrica alrededor de su media de lo que resulta que $n(\mu - x;\mu,\sigma) = n(\mu + x;\mu,\sigma)$, es decir, el valor de la densidad para la abscisa $\mu - x$ es igual al valor de densidad para la abscisa $\mu + x$. Por ejemplo, $P(\mu - \sigma < X < \mu) = P(\mu < X < \mu + \sigma) \approx 0.34$ y $P(\mu - \sigma < X < \mu + \sigma) \approx 0.68$. Al pie de la Figura 4.6 se pueden ver los porcentajes de área equivalentes a las probabilidades de que una variable aleatoria con distribución normal tome valores entre los puntos indicados. Por ejemplo, entre $\mu - \sigma$ y $\mu + \sigma$ está (aproximadamente) el 68% del área total bajo la curva lo que equivale a decir que la probabilidad de que la variable esté entre $\mu - \sigma$ y $\mu + \sigma$ es, aproximadamente, igual a 0.68 y la probabilidad de que la variable esté entre $\mu - 2\sigma$ y $\mu + 2\sigma$ es, aproximadamente, igual a 0.95. En la práctica, los cálculos de probabilidades asociadas con áreas bajo la curva de la distribución normal se realizan a partir de tablas o mediante programas de computadora de uso muy sencillo.

La función presenta su densidad máxima cuando la variable es igual a μ para luego ir decreciendo y acercándose asintóticamente al eje de abscisas sin cortarlo nunca. La distribución normal es, en realidad, una familia de distribuciones que difieren en su media y/o en su variancia. La representación gráfica de distribuciones normales con la misma variancia pero con distinta media se ve, como en la figura 4.7.

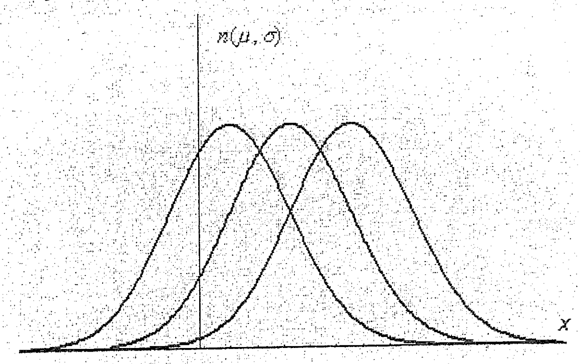


Figura 4.7. Distribuciones normales con igual σ^2 y diferente μ.

En cambio, la representación gráfica de distribuciones normales con la misma media pero con distintas variancias se ve como en la figura 4.8:

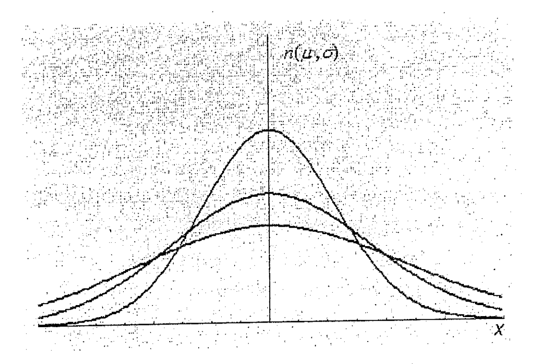


Figura 4.8. Distribuciones normales con igual μ y diferente σ^2.

La **distribución normal standard** (Z) es, simplemente, una distribución normal con media igual a 0 y variancia igual a 1 y sus probabilidades están extensivamente tabuladas. Dada la transformación de una variable normal (X) en normal standard (Z), la probabilidad acumulada correspondiente a un valor particular de X se puede leer fácilmente en una tabla de la distribución de Z puesto que:

$$\begin{aligned} N(x;\mu,\sigma) &= P(X \le x) \\ &= P[(\mu + \sigma \cdot Z) \le x] \\ &= P\left(Z \le \frac{x-\mu}{\sigma}\right) \\ &= N\left(\frac{x-\mu}{\sigma};0,1\right) \end{aligned}$$

Así que, dados $N(\mu,\sigma)$ y dos números reales cualesquiera x_1 y x_2, con $x_1 < x_2$, tendríamos:

$$\begin{aligned} P(a \le X \le b) &= N(b;\mu;\sigma) - N(a;\mu;\sigma) \\ &= N\left(\frac{b-\mu}{\sigma};0;1\right) - N\left(\frac{a-\mu}{\sigma};0;1\right) \end{aligned}$$

La representaciones gráficas de la distribución normal standard (a) y de su distribución de probabilidades acumuladas (b) son las representaciones en la figura 4.9:

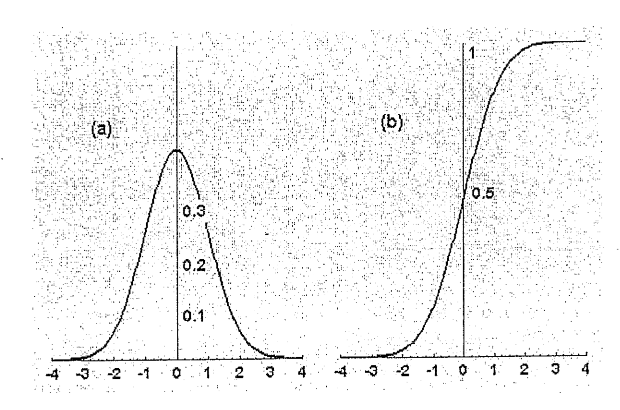


Figura 4.9. Distribución normal standard:

(a) Curva de densidad probabilidad.

(b) Distribución de probabilidad acumulada.

Ejemplo:

Una fábrica de objetos de aluminio produce cierto tipo de canal de aleación de aluminio. Se sabe que la rigidez de un canal producido por esta fábrica tomado al azar, medida en libras por pulgada^2 es una variable aleatoria con distribución normal con media μ = 2425 (lb/pulg^2) y σ = 115 (lb/pulg^2). Esta distribución se representa por la Figura 4.10.

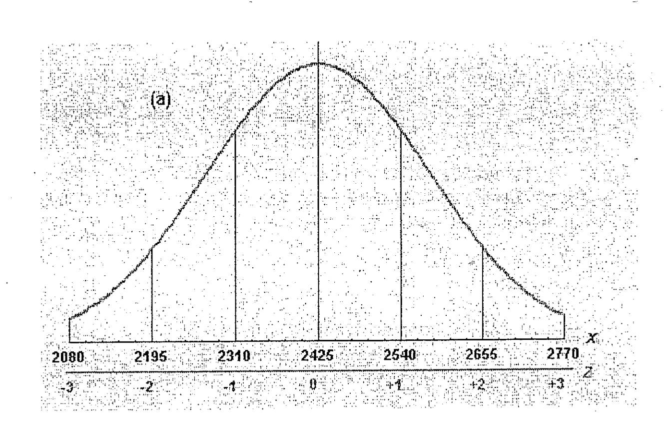


Figura 4.10.

Si se escoge al azar un canal de aleación de aluminio de este proceso:

(1) ¿cuál es la probabilidad de que tenga un valor de rigidez entre 2250 y 2425 lb/pulg^2?;

$$\begin{aligned} P(2250 \le X \le 2425) &= P(X \le 2425) - P(X \le 2250) \\ &= P\left(z < \frac{2425 - 2425}{115}\right) - P\left(z < \frac{2250 - 2425}{115}\right) \\ &\approx N(0) - N(11{,}52) \\ &= 0{,}5000 - 0{,}0643 \\ &= 0{,}4357 \end{aligned}$$

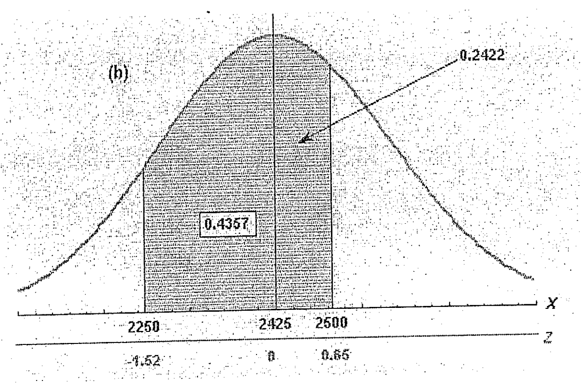
ver Figura 4.11

(2) ¿cuál es la probabilidad de que tenga un valor de rigidez entre 2250 y 2500 lb/pulg^2?;

$$\begin{aligned} P(2250 \le X \le 2500) &= P\left(z < \frac{2500 - 2425}{115}\right) - P\left(z < \frac{2250 - 2425}{115}\right) \\ &\approx N(0{,}65) - N(-1{,}52) \\ &= 0{,}7422 - 0{,}0643 \\ &= 0{,}6779 \end{aligned}$$

ver Figura 4.11

Figura 4.11.

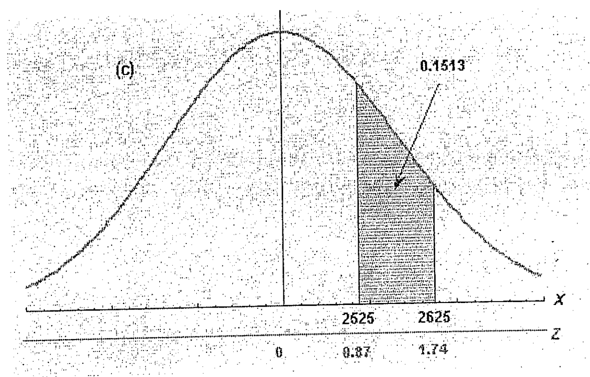


(3) ¿cuál es la probabilidad de que tenga un valor de rigidez entre 2525 y 2625 lb/pulg^2?;

$$\begin{aligned} P(2525 \le X \le 2625) &= P\left(z < \frac{2625 - 2425}{115}\right) - P\left(z < \frac{2525 - 2425}{115}\right) \\ &\approx N(1{,}74) - N(0{,}87) \\ &= 0{,}9591 - 0{,}8078 \\ &= 0{,}1513 \end{aligned}$$

ver Figura 4.12

Figura 4.12.



(4) ¿cuál es la probabilidad de que tenga un valor de rigidez mayor de 2500 lb/pulg^2?;

$$P(X > 2500) = 1 - P(X < 2500)$$

$$= 1 - P\left(z < \frac{2500 - 2425}{115}\right)$$

$$\approx 1 - N(0{,}65)$$

$$= 1 - 0{,}7422$$

$$= 0{,}2578$$

ver Figura 4.13

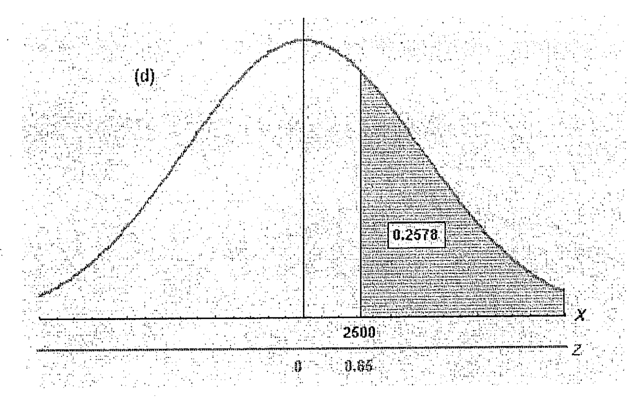


Figura 4.13

(5) ¿cuál es la probabilidad de que tenga un valor de rigidez menor de 2200 lb/pulg^2?

$$P(X < 2200) = P\left(z < \frac{2200 - 2425}{115}\right)$$

$$\approx N(-1{,}96)$$

$$= 0{,}025$$

ver Figura 4.14

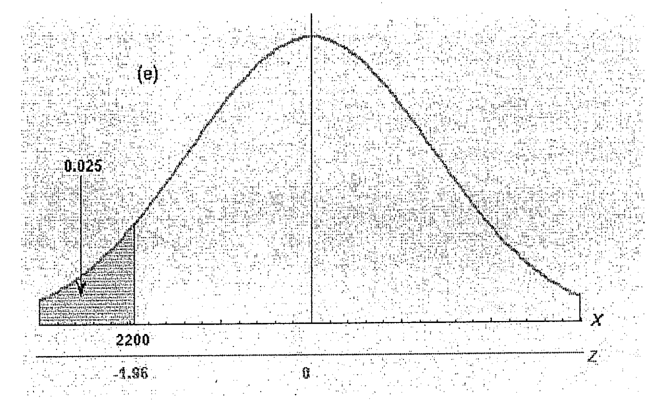


Figura 4.14..

A continuación presentaremos dos distribuciones que se emplean para el cálculo de probabilidades en situaciones especiales que veremos algunas clases más adelante.

La distribución χ^2

Si $Z_1, Z_2, \cdots, Z_\nu$ son variables normales standard independientes, la suma de sus cuadrados se dice que es una variable χ^2 (léase ji cuadrado) con ν grados de libertad. Es decir:

$$\chi_\nu^2 = Z_1^2 + Z_1^2 + \cdots + Z_\nu^2 \tag{4.16}$$

El concepto de **grados de libertad** es un concepto del álgebra de espacios vectoriales. Es el nombre dado al número de observaciones inicialmente independientes que hay en una suma de cuadrados. No discutiremos aquí la base teórica de este concepto sino que lo abordaremos heurísticamente.

El parámetro ν define a la distribución χ^2 y hay una distribución χ^2 para cada valor de ν, como puede verse en la figura 4.15:

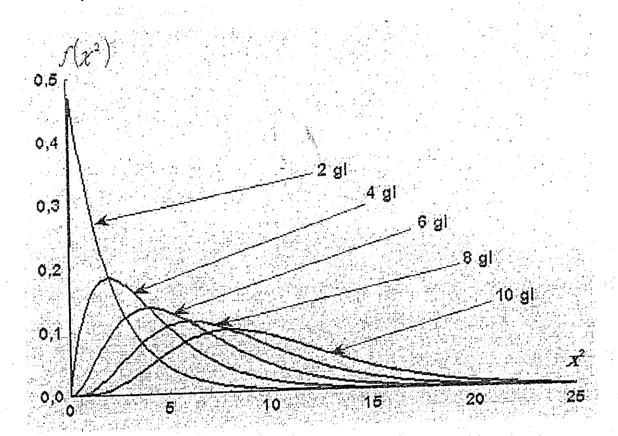


Figura 4.15. Familia de distribuciones χ^2 con diferentes grados de libertad.

Las tablas de la distribución χ^2 presentan los valores de χ^2 para algunas probabilidades específicas (ver Tabla en la página 128 y el menú ***Probabilidades y Cuantitles*** de ***Infostat***). Veamos dos ejemplos de utilización de las tablas para χ_{15}^2:

$$P(X > 7.26) = P(7.26 < \chi_{15}^2 < \infty)$$
$$= 0.95 \text{ y } P(X > 5.23)$$
$$= P(5.23 < \chi_{15}^2 < \infty)$$
$$= 0.99.$$

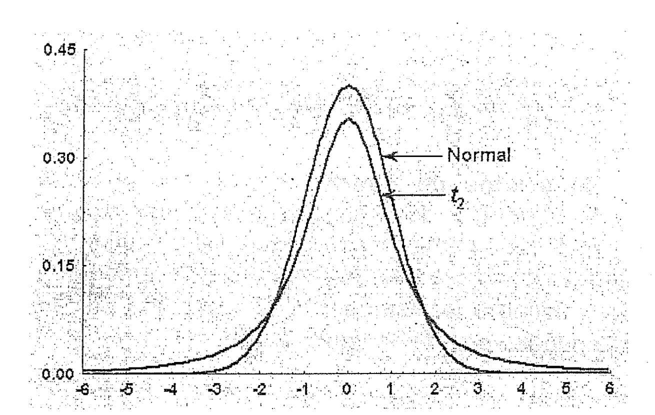
La distribución *t* de Student

Una distribución ***t* de Student** es la distribución de probabilidad de una variable aleatoria que resulta de dividir una variable con distribución normal standard por la raíz cuadrada de una otra con distribución χ^2 dividida por sus grados de libertad:

$$t_{n-1} = \frac{Z_0}{\sqrt{\frac{1}{n-1}\cdot\left(Z_1^2 + Z_2^2 + \cdots + Z_{n-1}^2\right)}} = \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} \tag{4.17}$$

donde $Z_0, Z_1, Z_2, \ldots, Z_n$, son $n + 1$ variables normales standard independientes. Esta es una distribución *t* de Student con $n - 1$ grados de libertad. (En la página 127 se presenta la tabla de esta distribución. Ver, también, el menú ***Probabilidades y Cuantitles*** de ***Infostat***).

En la figura 4.16 se representan una distribución t de Student y una distribución normal con finales comparativos:



4.16. Comparación entre las funciones de densidad de las distribuciones normal standard y t de Student con 2 grados de libertad

Ejercicios

4.1. Un inversor que dispone de $100.000 para realizar una inversión tiene dos alternativas. La primera es colocar el dinero en un plan de inversión con una rentabilidad anual fija del 15%. La segunda alternativa es colocar el dinero en otro plan de inversión cuya rentabilidad anual varía entre el 5 y el 30 % según las condiciones económicas que prevalezcan. La historia de este último tipo de inversión permite suponer que la distribución de probabilidad de sus valores de rentabilidad es la que figura en la tabla:

Rentabilidad anual (%)	30	25	20	15	10	5
Probabilidad	0,08	0,26	0,34	0,23	0,08	0,01

Si este supuesto fuera correcto:

a. *¿Que probabilidad habría de obtener mayor rentabilidad con el segundo plan que con el primero?* 0,68

b. *¿Cuál sería la rentabilidad esperada con el segundo plan?* 20

c. *¿Cual sería el desvío standard de la rentabilidad del segundo plan?* 9,35

d. *A partir de los resultados anteriores, explicar cuál plan le conviene elegir al inversor.*

4.2. La probabilidad de infección con *oídio* (enfermedad fúngica) en plantas de zapallito redondo en las quintas del cinturón hortícola del Gran Buenos Aires es 0.15. Si usted es contratado por el Ministerio de Asuntos Agrarios de la provincia de Buenos Aires para elaborar un informe acerca del estado de la enfermedad en dicha área y decide visitar 15 quintas, ¿cuál es la probabilidad esperada para los siguientes sucesos:

a. *A lo sumo 3 quintas presenten cultivos infectados.*
b. *Sólo 5 quintas presenten cultivos infectados.*
c. *Al menos 4 quintas presenten cultivos infectados.*

4.3. Se denomina Poder Germinativo a la proporción de semillas de un lote que germinan cuando se las coloca en condiciones apropiadas de tempe-

ratura y humedad. La etiqueta de una bolsa de semillas dice que su poder germinativo es de 95%. Para evaluar la veracidad de esta especificación se toman de la bolsa 10 semillas al azar se4 las coloca durante 10 días en condiciones apropiadas para la germinación. Al cabo de ese período se cuenta y registra el número de semillas que germinaron:

a. *¿De qué tipo es la variable aleatoria registrada?* Cuantitativo Discret.
b. *¿Qué valores puede tomar dicha variable aleatoria?* 0 a 10

En caso de que lo que dice la etiqueta fuera cierto:

c. *¿Cuál sería la probabilidad de que, en la prueba descripta, germinaran todas las semillas?* 0,60
d. *¿Cuál sería la probabilidad de que, en la prueba descripta, quedaran 2 semillas sin germinar?* 0,07
e. *¿Cuál sería la probabilidad de que, en la prueba descripta, quedaran 2 o más sin germinar?* 0,086
f. *¿Cuánto valdrían la esperanza y la variancia del número de semillas germinadas* E(n)=9,5 Var(n)=0,475

4.4. Existen insectos como el Tatadios (*Mantis religiosa*) que son considerados útiles para la agricultura porque se alimentan de otros insectos que dañan a los cultivos. Si cuando un Tatadios encuentra un insecto presa tiene una probabilidad de capturarlo de 0,25

g. *¿Cuántos insectos debe encontrar para que la probabilidad de que capture al menos uno sea se 0,8?*

Si encontrara esa cantidad de insectos presa por día:

h. *¿Cuál sería el número esperado insectos que captura por día?*
i. *¿Cuánto variaría el número de insectos que captura por día?*

4.5. La duración de la vida de las plantas del pasto bianual *Bromus unioloides* es una variable aleatoria X que puede tomar valores ente 0 y 2 años. Si la función de densidad de probabilidad de dicha variable aleatoria fuera:

$$f(x) = 1 - \frac{x}{2}, \text{ para } 0 \leq x \leq 2.$$

a. *Graficar la función f(x).*
b. *Verificar que f(x) es una función de densidad.*
c. *Calcular la función de distribución de probabilidades*
d. *Calcular la probabilidad de que una planta de* Bromus unioloides *tomada al azar viva menos que un año*
e. *Calcular la probabilidad de que una planta de* Bromus unioloides *tomada al azar viva más que un año y medio*
f. *Calcular la probabilidad de que una planta de* Bromus unioloides *tomada al azar viva entre una año y un año y medio?*
g. *Verificar que las tres probabilidades calculadas suman 1 y explicar por qué,*
h. *Calcular la esperanza matemática y la variancia de la duración de la vida de una planta de* Bromus unioloides *tomada al azar.*

4.6. El peso de los terneros de raza Aberdeen Angus recién nacidos es una variable aleatoria con distribución aproximadamente normal con media de 32 kg y varianza de 6,25 kg2.

a. *¿Cuál es la probabilidad de que un ternero de raza Aberdeen Angus recién nacido tomado al azar pese entre 27 kg y 37 kg?*
b. *¿Cuál es la probabilidad de que un ternero de raza Aberdeen Angus recién nacido tomado al azar pese más de 39 kg?*
c. *¿Cuál es el peso que deja por debajo al 90% de los pesos de todos los terneros Aberdeen Angus recién nacidos?*
d. *¿Cual es la probabilidad de que entre dos terneros de raza Aberdeen Angus recién nacidos tomados al azar uno pese más de 32 kg y el otro menos de 32 kg?*

4.8. En un área de la provincia de La Pampa, el 25% de los establecimientos han incorporado especies forrajeras mejoradas en sus pastizales naturales mediante intersiembra. En dichos establecimientos, la duración de la invernada (engorde de los novillos para faena) es una variable aleatoria con distribución aproximadamente Normal, con media de 650 días y desviación estándar de 45 días. En cambio, en los establecimientos restantes, la duración de la invernada es una variable aleatoria con distribución aproximadamente Normal, con media de 770 días y desviación estándar de 85 días.

a. *¿Cuál es la probabilidad de que en una muestra de 25 establecimientos de esta área tomados al azar, 5 o menos hayan incorporado especies forrajeras mejoradas?*
b. *¿Cuál es la probabilidad de que la duración de la invernada se prolongue más de 770 días en los establecimientos que han incorporado especies forrajeras mejoradas?*
c. *¿Cual es la probabilidad de que la duración de la invernada sea menor que 650 días en los establecimientos que no han incorporado especies forrajeras mejoradas?*

4.6. El 40% de los animales de un rodeo son de raza *A* y el resto, de raza *B*. Si el peso de los animales de la raza *A* sigue una distribución normal con media 250 kg y varianza 400 kg^2 y el peso de los animales de la raza *B* sigue una distribución aproximadamente normal con media 270 kg y desvío típico 30 kg:
¿Qué porcentaje de animales tiene peso superior a 240 kg?

4.10. En un área del oeste de la Región Pampeana, se ha determinado que la sequía es el principal factor que afecta la seguridad de cosecha de cereales de invierno como el trigo y el centeno. En esta área, el total de lluvias inverno-primaverales es una variable aleatoria con distribución aproximadamente normal con media igual a 300 mm y desvío standard igual a 100 mm. Cuando durante el período inverno-primaveral llueven menos de de 250 mm se compromete seriamente la cosecha de trigo, en cambio el cultivo de centeno, más resistente a la sequía, produce mientras llueva más de 200 mm

a. *¿Cuál es la seguridad de cosecha de trigo en esta área?*
b. *¿Cuál es la seguridad de cosecha de centeno?*
c. *¿Cual es la probabilidad de que se pierda la cosecha de trigo pero no la de centeno?*
d. *¿Cuál es la probabilidad de que no se pierda ninguna de las dos cosechas?*

DISTRIBUCIONES POR MUESTREO

El procedimiento estadístico de extracción de información útil es una secuencia que comienza con la obtención de una **muestra aleatoria** de n unidades muestrales tomadas al azar de una **población** de tamaño N, continúa con el registro de los valores que toma una **variable aleatoria** en cada una de las unidades muestrales y culmina con la aplicación de la Teoría de Probabilidades para realizar una afirmación acerca de los valores de uno o más **parámetros** de la distribución de probabilidad de la variables aleatoria en la población. Esta última afirmación es conocida como **inferencia estadística** y es realizada a partir de funciones de los valores muestrales de la variable aleatoria denominadas genéricamente **estadísticas**.

Las estadísticas muestrales son entonces funciones de las los valores registrados de la variable aleatoria. Un ejemplo de una estadística es la media aritmética de los valores registrados en la muestra. Tanto la media aritmética muestral como cualquier otra estadística es por lo tanto una función de variables aleatorias, por ello **es también una variable aleatoria**. Su valor varía de muestra en muestra de modo que, antes de obtener la muestra, hay incertidumbre acerca de qué valor tomará exactamente la estadística. Como cualquier variable aleatoria, cada estadística tiene una distribución de probabilidad. Para poder hacer inferencia estadística resulta necesario conocer la distribución de probabilidades de las estadísticas utilizadas. La distribución de una estadística particular depende de (a) la distribución de probabilidad de la variable aleatoria registrada en la población, (b) del tamaño de la muestra aleatoria.

Para que la muestra sea realmente una **muestra aleatoria** es necesario que contenga un conjunto de n unidades muestrales extraídos de la población utilizando un procedimiento de sorteo que otorgue igual probabilidad de entrar en dicho conjunto a todas las unidades de la población. De este modo, las realizaciones de la variable aleatoria X_1, X_2, ..., X_n registradas en cada una de las unidades muestrales extraídas (ai) son todas independientes entre ellas y (b) provienen de la misma distribución de probabilidades. Estas condiciones son necesarias para que los estadísticos calculados tengan las distribuciones de probabilidad que presentamos aquí.

La media muestral y la variancia muestral

Entre los distintos estadísticos que se pueden calcular a partir de los datos contenidos en una muestra, hay dos que resaltan por su importancia y sus aplicaciones: la **media muestral** y la llamada **variancia muestral**. Tal como la hemos definido el capítulo 2, la media muestral ($\bar{x}$) y la así llamada variancia muestral (s_{n-1}^2) se calculan como:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ y} \qquad (5.1)$$

$$s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (5.2)$$

La razón por la cual son tan importantes estos dos estadísticos es que sirven para estimar la media y la variancia de la variable aleatoria estudiada en la población total. Ahora, un estadístico, dado que no es otra cosa que una cantidad que se calcula a partir de los datos de una muestra, es, como los datos de la muestra, una variable aleatoria. Entonces, cuando se tomen muchas muestras, mostrarán la variación propia de una variable aleatoria de muestra en muestra. Así que se deben conocer las propiedades de estas dos variables aleatorias tan importantes, es decir, cómo se espera que sea su comportamiento al extraer

muestras y, sobre todo, cómo cambiarán sus propiedades, al cambiar el tamaño de las muestras que se tomen. Por ejemplo, una propiedad fundamental de la media muestral es su esperanza. Es muy fácil deducir que si la esperanza de una variable aleatoria x es igual a μ, entonces la esperanza de su media muestral será, también: $E(\bar{x}) = \mu$. Como se dijo antes, en el muestreo de una población, la esperanza de todas las medias muestrales que se pueden calcular es igual a la media poblacional (μ). Pero se debe considerar que dichas medias muestrales mostrarán una variación de muestra en muestra, es decir, entre todos los valores posibles que la media muestral puede tomar: esa es la **variancia de la media muestral,** $\sigma^2(\bar{x})$. Dado que la variancia de la variable aleatoria x en la población es igual a σ^2, la variancia de la media muestral es igual a $\sigma^2(\bar{x}) = E\left[(\bar{x}-\mu)^2\right]$ y es fácil demostrar que $\sigma^2(\bar{x}) = \dfrac{\sigma^2}{n}$, que es una propiedad muy importante de la variancia de la media muestral puesto que está indicando que la distribución de la media muestral se concentra cada vez más en el entorno de μ, a medida que aumenta el tamaño de la muestra (n). Esto es lo mismo que decir que, cuanto mayor sea el tamaño muestral, más confianza se podrá tener en que la media de la muestra estará más cerca de la media poblacional desconocida (μ).

Generación de la distribución por muestreo de una estadística

Veremos un ejemplo de cómo se puede generar la distribución por muestreo de una estadística. Supongamos que una distribuidora de bebidas vende un refresco en 3 tamaños de botella: 500 cm^3, 750 cm^3 y 900 cm^3. El 50% de los refrescos que vende son de 500 cm^3, el 30%, de 750 cm^3 y el 20% restante de 900 cm^3. En un puesto de venta aparecen 2 clientes. Sea X_1 el tamaño de botella que compra el primer cliente y X_2 el tamaño de botella que compra el segundo cliente y supongamos que X_1 y X_2 son independientes, es decir, suponemos que la compra realizada por el primer cliente no influye para nada en la compra que habrá de hacer el segundo cliente. Tanto X_1 como X_2 tiene la distribución de probabilidad que se mencionó antes, es decir:

Cuadro 5.1

x	500	750	900
$P(x)$	0.50	0.30	0.20

Así que los dos clientes constituyen una muestral aleatoria de esta distribución de probabilidades. La siguiente tabla enumera todos los posibles pares de valores de X_1 y X_2 con sus respectivas probabilidades calculadas bajo el supuesto de independencia y los valores de media ($\bar{X}$) resultantes.

Cuadro 5.2

x_1	x_2	$p(x_1;x_2)$	$\bar{x}$
500	500	0.25	500
500	750	0.15	625
500	900	0.10	700
750	500	0.15	625
750	750	0.09	750
750	900	0.06	825
900	500	0.10	700
900	750	0.06	825
900	900	0.04	900

Por tanto, la distribución por muestreo de $\bar{X}$ es:

Cuadro 5.3

$\bar{X}$	500	625	700	750	825	900
$P(\bar{X})$	0.25	0.30	0.20	0.09	0.12	0.04

Tanto la media de la distribución original como la media de la distribución de $\bar{X}$ son iguales a 655, confirmando que $E(\bar{X}) = \mu$. La variancia de la distribución original es 26725 mientras que la variancia de la distribución de $\bar{X}$ es igual a 13362.5, confirmando que $\sigma^2(\bar{X}) = (\sigma^2/n)$. En nuestro ejemplo $n = 2$ así que 13362.5 = 26725/2. Además, vemos que la distribución de probabilidad de $\bar{X}$ es diferente de la de X, primero porque como vimos recién, $\bar{X}$ tiene menor variancia que X y también porque la probabilidad está algo más concentrada en los valores cercanos a μ., la media poblacional. Para muestras de mayor tamaño, estas características son más acentuadas.

La relación entre el tamaño las muestras y la distribución de probabilidad de la media muestral es definida por el teorema más importante de la estadística, denominado **Teorema Central del Límite** cuyo enunciado se presenta a continuación. Este teorema es fundamental para desarrollar todas las herramientas de inferencia estadística que veremos más adelante para, por ejemplo, estimar la media poblacional de una variable aleatoria con una precisión deseada y conocida.

El Teorema Central del Límite

El **Teorema Central de Límite** (TCL) en palabras, dice que *si una población tiene una media μ y variancia σ^2, finitas, entonces, a medida que el tamaño de la muestra (n) aumenta, la distribución de la media de la muestra ($\bar{x}$), tiende a la distribución normal con media μ y variancia* $\frac{\sigma^2}{n}$. En términos de la distribución normal standard:

$$P(\bar{x} \le \bar{x}_0) = N\left(\frac{\bar{x}_0 - \mu}{\sigma/\sqrt{n}}\right) \qquad (5.3)$$

donde $\bar{x}_0$ es un valor particular de $\bar{x}$.

La precisión de esta probabilidad depende del tamaño de la muestra y de la distribución de la variable aleatoria *X*. Si *X* tiene distribución normal, las probabilidades serán exactas, sin importar cuán pequeña sea la muestra. Si no se conoce la distribución de *X*, la probabilidad será más exacta cuanto mayor sea *n*.

Ejemplo.
Una empresa produce bolsas de un producto agroquímico con un peso medio de 50 kg y una variancia de 4 kg^2. Se toma una muestra de 100 bolsas. Asumiendo que los pesos de las bolsas son independientes, según el TCL, el peso medio de una muestra, *M*, debería distribuirse de manera aproximadamente normal así que, podemos calcular probabilidades. Por ejemplo;

$$P(M < 49.7) = P\left(z < \frac{49.7 - 50}{2/\sqrt{100}}\right)$$
$$\cong P(z < -1.5)$$
$$\cong 0.0668$$

$$P(M > 50.4) = 1 - P\left(z < \frac{50.4 - 50}{2/\sqrt{100}}\right)$$
$$\cong 1 - P(z < +2.0)$$
$$\cong 1 - 0.9773$$
$$= 0.0227$$

$$P(49.8 \le M \le 50.6) = P\left(z < \frac{50.6 - 50}{2/\sqrt{100}}\right) - P\left(z < \frac{49.8 - 50}{2/\sqrt{100}}\right).$$
$$\cong 0.84000$$

A continuación, aplicando el TCL, veremos cómo es la distribución de algunos estadísticos de uso muy común, cuando se efectúan muestreos sobre distintos tipos de poblaciones.

Distribución por muestreo de la media

Hemos visto, al principio del capítulo, que la media muestral, $\bar{x}$, tiene esperanza igual a μ y variancia igual a (σ^2/n) y, por ende, error standard igual a $(\sigma/\sqrt{n})$ y que cuando $n \to \infty$, $\sigma_{\bar{x}} \to 0$. El TCL establece que, cuando n es grande, la función de distribución acumulativa de x es:

$$P(\bar{x} \le \bar{x}_0) \cong N\left(\frac{\bar{x}_0 - \mu}{\sigma/\sqrt{n}}\right) \tag{5.4}$$

Es decir que $\dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ es una variable aleatoria con distribución normal standard.

Distribución por muestreo de la diferencia entre dos medias (muestras independientes)

Asimismo, más allá del interés en estimar la probabilidad de una media muestral determinada, muchas veces lo que interesa realmente es la diferencia entre dos medias muestrales, o sea, la comparación de dos medias muestrales. Dadas dos muestras tomadas independientemente una de la otra (de dos poblaciones con medias μ_1 y μ_2), con tamaños muestrales n_1 y n_2, con medias $\bar{x}_1$ y $\bar{x}_2$, nos interesa utilizar la diferencia entre las medias muestrales, $\Delta\bar{x} = \bar{x}_1 - \bar{x}_2$ para estimar la verdadera diferencia entre los parámetros poblacionales, es decir, entre μ_1 y μ_2, $\Delta\mu = \mu_1 - \mu_2$. Según el TCL, la distribución por muestreo de $\Delta\bar{x}$ se aproxima a una

distribución normal con media $\Delta\mu$ y error standard $\sigma(\Delta_{\bar{x}}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, (donde σ_1^2 y σ_2^2 son las variancias de la variable de interés en las dos poblaciones, respectivamente) cuando n_1 y n_2 son grandes. Así que la probabilidad de una dada diferencia puede aproximarse mediante la expresión:

$$P(\Delta\bar{x} \leq \Delta\bar{x}_0) \cong N\left(\frac{\Delta\bar{x}_0 - \Delta\mu}{\sigma(\Delta\bar{x})}\right) \tag{5.5}$$

Ejemplo.
El rendimiento medio en [Kg/Ha] de maíz en la localidad A es de 4700 con una variancia de 47000 [Kg/ha]^2 y en la localidad B, es de 4200 [Kg/Ha] con una variancia de 100000 [Kg/Ha]^2. Si se eligen al azar 49 establecimientos de la localidad A y 80 de la localidad B y se determinan sus rendimientos medios de maíz, ¿cuál es la probabilidad de que el rendimiento medio de la muestra A sea por lo menos 550 [Kg/Ha] mayor que el de la muestra B?

$$\begin{aligned}\Delta\mu &= 4700 - 4200\\ &= 500 \text{ [Kg/Ha] y}\end{aligned}$$

$$\begin{aligned}\sigma(\Delta_{\bar{x}}) &= \sqrt{\frac{47000}{49} + \frac{100000}{80}}\\ &\cong 47.00\end{aligned}$$

La probabilidad buscada es:

$$\begin{aligned}P(\Delta\bar{x} \geq 550) &\cong 1 - N\left(\frac{550 - 500}{47}\right)\\ &\cong 1 - N(1.064) \quad \text{y}\\ &\cong 0.1446\end{aligned}$$

Distribución por muestreo de la variancia muestral

Si la variable aleatoria x tiene distribución normal en la población, entonces la distribución por muestreo de la variancia muestral, s_{n-1}^2, puede obtenerse de: $s_{n-1}^2 = \chi_{n-1}^2 \cdot \frac{\sigma^2}{n-1}$, es decir que el estadístico muestral que tiene distribución χ^2 es $\frac{(n-1)\cdot s_{n-1}^2}{\sigma^2}$, donde χ_{n-1}^2 es una distribución χ^2 con $n-1$ grados de libertad y σ^2 es la variancia de x en la población.

Distribución por muestreo de $\frac{\bar{x}-\mu}{s/\sqrt{n}}$

Cuando no se conoce σ^2, ya no se puede utilizar la variable $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ que, según el TCL tiene distribución normal estándar. En ese caso, se usa $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ que tiene distribución t de Student que presentamos en el capítulo 4. Dada una muestra de tamaño n:

$$P(\bar{x} < \bar{x}_0) \approx P\left(t_{n-1} < \frac{x_0 - \mu}{s/\sqrt{n}} \right) \qquad (5.6)$$

donde s es el desvío standard muestral y t_{n-1} es una variable t de Student con n – 1 grados de libertad.

Ejercicios

5.1 Sea la variable aleatoria X cuya distribución de probabilidad es la siguiente:

x	1	2	3	8
$P(X=x)$	0.1	0.4	0.4	0.1

a. *Graficar la distribución de probabilidades de X.*

b. *Calcular μ la esperanza de X y σ^2, la variancia de X.*

c. *Hallar la distribución por muestreo de $\overline{X}_2$ la media muestral de una muestra aleatoria de tamaño n = 2. Para ello, determinar todos los valores que puede tomar $\overline{X}_2$ y encontrar la probabilidad asociada con cada uno de ellos. Graficar y comparar con la distribución producida en a.*

d. *Usar el resultado encontrado en (b) para obtener $\mu_{\overline{X}_2}$ y $\sigma^2_{\overline{X}_2}$.*

e. *¿Que relaciones se verifican entre μ y $\mu_{\overline{X}_2}$ y entre σ^2 y $\sigma^2_{\overline{X}_2}$?*

5.2 En una población de plantas de cebada, hay dos genotipos claramente distinguibles por la cantidad de hileras de granos en sus espigas: uno de ellos tiene 2 hileras y el otro, 6 hileras. La población está compuesta por un 70% de plantas de 2 hileras de granos y un 30% de plantas de 6 hileras de granos.

a. *Calcular la esperanza y la variancia del número de hileras de granos por planta en esta población.*

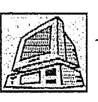
b. *Detallar las 8 diferentes composiciones posibles de las muestras aleatorias de 3 plantas obtenidas de esta población.*

c. *Para cada una, calcular su probabilidad y la media muestral de los el números de hileras de granos.*

d. *Calcular la esperanza y la variancia de las medias muestrales obtenidas en el punto (c).*

e. *¿Qué relaciones se observan entre los parámetros calculados en el punto (d) y los calculados en el punto (a)?*

5.3 Suponer que una muestra aleatoria de tamaño n = 25, es seleccionada de una población con media μ, y desvío estándar σ. Para cada uno de los siguientes valores de μ y σ, determinar los valores de $\mu_{\bar{x}}$ y $\sigma_{\bar{x}}$:
(a) μ = 10 y σ = 3; (b) μ = 100 y σ = 25; (c) μ = 20 y σ = 40; (d) μ = 10 y σ = 100.



5.4 Si el peso de los novillos de raza Aberdeen Angus de 36 meses de edad es una variable aleatoria con distribución normal con valor esperado de 420 kg y variancia 64 kg2:

a. *¿Cuál es la probabilidad de que un novillo tomado al azar pese a lo sumo 425 kg?* 0,7324
b. *¿Cuál es la probabilidad de que un novillo tomado al azar pese exactamente 420 kg?* 0
c. *¿Cuál es la probabilidad de que dos o más entre cuatro novillos tomados al azar pesen a lo sumo 425 kg?* 0,9372
d. *¿Cuál es la probabilidad de que la media de los pesos de cuatro novillos tomados al azar sea a lo sumo 425 kg?* 0,8944
e. *¿Qué distribución de probabilidad aproximada tienen las medias aritméticas de los pesos de 4 novillos tomados al azar?* Normal

5.5 Los barros cloacales son enmiendas orgánicas que se utilizan para mejorar la productividad de forraje en algunos pastizales. Sin embargo, son ricos en metales pesados tóxicos como el plomo que puede ser absorbido por las plantas y acumularse luego en la carne de los animales que las comen. Con fines bromatológicos, la carne es clasificada en las siguientes tres categorías según su contenido de plomo en partes por millón (ppm = mg / kg) :

	Inofensiva	Levemente Tóxica	Tóxica
Concentración de Pb (ppm)	< 0.1	[0.1-0.5)	≥ 0.5

Si la concentración de plomo en la carne de los terneros producidos en lotes tratados con barros cloacales es una variable aleatoria con distribución aproximadamente normal con media poblacional $\mu = 0.2$ ppm y varianza $\sigma^2 = 0.02$ ppm.

a. *¿Cual es la probabilidad de que la carne de un ternero tomado al azar en un establecimiento que aplica barros cloacales resulte clasificada como Levemente Tóxica o Tóxica?*
b. *¿Cuál es la probabilidad de la carne de dos o más entre nueve terneros tomados al azar resulte clasificada como Levemente Tóxica o Tóxica?*
c. *¿Cuál es la probabilidad de que la media aritmética de la concentración de plomo en la carne de una muestra de 9 terneros tomados al azar corresponda a las categorías Levemente Tóxica o Tóxica?*

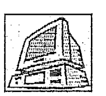
5.6 En una región semi-desértica donde llueven 200 mm/año es razonable suponer que la productividad primaria neta anual (PPNA) promedio de los pastizales es de 86 g/m² con un desvío estandar de 40 g/m². Consideremos una muestra cualquiera de 40 pastizales tomados al azar dentro de dicha región cuyos valores de PPNA son y_i, $(i = 1, \ldots, 40)$ y definamos los siguientes estadísticos:

$$y = \frac{1}{40}\sum_{i=1}^{40} y_i \qquad S^2 = \frac{1}{39}\sum_{i=1}^{40}(y_i - y)^2$$

a. *¿Cuál es el valor esperado de $\bar{y}$?*
b. *¿Cuál es la varianza de $\bar{y}$?*
c. *¿Qué distribución de probabilidad aproximada tiene y ?*
d. *¿Cuál es la probabilidad que y supere los 90 g/m ^{2}?*
e. *¿Cuál es la probabilidad de que 2 entre 3 muestras de 40 pastizales tomados al azar tengan valores de $\bar{y}$ que superen los 90 g/m ^{2}?*
f. *¿Cuál es el valor esperado de s^2?*
g. *¿Qué distribución de probabilidad aproximada tiene del estadístico*

$$\frac{\bar{y}-86}{\frac{s}{\sqrt{40}}} ?$$

5.7 El siguiente conjunto de datos representa un censo efectuado sobre el tamaño de manzanas en una línea de empaque del Alto Valle del Río Negro durante la época de cosecha. De acuerdo a estos datos, el tamaño de las manzanas (diámetro en milímetros) sigue una distribución Normal con μ = 78 mm y σ = 4 mm. En base a esta tabla, escoger 10 muestras aleatorias de n =10 y probar la distribución de los estimadores y su relación con los parámetros.



	1	2	3	4	5	6	7	8	9	10
1	80.1	81.0	76.1	77.9	73.3	70.5	81.4	76.9	72.1	78.0
2	72.9	81.2	82.5	71.4	71.8	73.5	76.7	78.8	79.2	77.4
3	74.6	79.3	76.2	79.5	78.2	73.9	84.6	75.3	82.2	72.4
4	77.6	79.7	78.7	84.2	85.5	79.4	79.1	77.1	82.7	74.0
5	78.4	77.0	76.4	80.2	68.7	76.5	81.1	74.5	73.7	75.0
6	79.8	81.7	81.8	83.1	75.6	75.8	76.6	78.5	74.2	75.2
7	75.7	85.0	83.6	86.2	76.8	75.9	83.4	80.6	77.8	80.0
8	77.5	78.9	87.3	75.4	77.7	76.0	69.8	80.7	81.5	78.1
9	82.3	78.3	77.2	79.9	73.1	77.3	74.9	83.9	74.3	79.0
10	74.8	78.6	72.6	80.8	80.5	80.3	82.0	71.0	82.9	81.7

5.8. El rendimiento promedio de los cultivos de un híbrido de maíz en la región de la Pampa Ondulada es de 10 tn/ha y que el desvío estándar es de 1,5 tn/ha.

a. *Identificar a la población, a las unidades muestrales y a la variable aleatoria a las que se hace referencia.*

Supongamos que se tomarán de la región 25 cultivos de de dicho híbrido elegidos al azar y se calculará la media aritmética de sus rendimientos:

b. *Explicar por qué la media aritmética es una variable aleatoria. ¿Cuál es la población correspondiente?*
c. *¿Qué distribución de probabilidad aproximada tiene la media aritmética en cuestión?*
d. *¿Cuál es la probabilidad de que dicha media aritmética supere los 10.500 kg/ha?*
e. *¿Cuál es la probabilidad de que, entre 3 muestras aleatorias como la referida, dos muestras tengan media aritmética de los rendimientos mayor que 10.500 kg/ha de materia seca/m ^{2}?*

5.9. Una compañía envasadora de harina afirma que los paquetes que produce tienen un peso promedio de 1000 g y que la variancia de los pesos es de 36 g^2.

Si la afirmación de la compañía fuese cierta:

a. *¿Cuál sería la distribución de probabilidad aproximada de la media aritmética de los pesos de 36 paquetes tomados al azar?*

b. *¿Cuál sería la probabilidad de que la media aritmética de los pesos de 36 paquetes tomados al azar estuviese comprendida entre 998 y 1002 g?*

Teniendo en cuenta las respuestas a los puntos anteriores:

c. *Discutir en qué medida se puede dar crédito a la afirmación de la compañía si se encuentra que la media aritmética de los pesos de 36 paquetes tomados al azar es de 998 g.*

ESTIMACIÓN DE PARÁMETROS

En un estudio acerca de la disponibilidad de alimento para la dieta de elefantes marinos en la Península de Valdés, se necesita determinar la biomasa promedio de las presas disponibles de una determinada especie. Obviamente es imposible pesar a todas las presas de esa especie que se encuentran en el espacio que los elefantes marinos pueden explorar en la plataforma continental en una temporada. En cambio, se puede diseñar un muestreo aleatorio que abarque el área de distribución de elefantes marinos en el mar (descripta por seguimiento satelital en campañas anteriores), capturar en cada sitio una presa de la especie en cuestión y pesarla. El conjunto de las capturas tomadas al azar constituye una **muestra aleatoria**, representativa de la **población** formada por todos los animales que hubieran podido ser capturados. El peso es una **variable aleatoria** que puede tomar diferentes valores según cual sea el animal capturado.

Nuestra intención al tomar una muestra es la de hacer una **inferencia**. Este término lo usamos en Estadística para denominar al procedimiento con el que hacemos afirmaciones acerca de **parámetros** de la población mediante los números que observamos en la muestra. En el caso del estudio sobre la dieta de los elefantes marinos, el parámetro sobre el cual se hace inferencia es el peso promedio de todas las presas de la población. Para hacer esta inferencia, es fundamental que cualquier individuo de la población de interés haya tenido igual probabilidad de entrar en la muestra. En ese caso, la muestra es representativa de la población. Una muestra aleatoria formada por *n* unidades de observación provee una colección de *n* valores (realizaciones) de la variable aleatoria. Estas realizaciones (a) son independientes y (b) provienen de la misma distribución de probabilidad.

Para tener una idea del valor del parámetro que desconocemos tomamos una **muestra** de los pesos de las presas. Supongamos que son 100 presas en la muestra. Con una balanza de la precisión adecuada y con mucho cuidado, medimos los pesos de las 100 presas de la muestra y calculamos su promedio. ¿Qué nos dice el valor de la media de la muestra acerca de la media de la población? Por un lado, definitivamente no esperamos que el valor de la media de la muestra coincida exactamente con el de la población. Por otra parte, no tenemos mejor información respecto a la media de la población que la que extraigamos de la muestra. Por último, sería muy extraño que si la población de presas tiene, por decir algo, un peso promedio de 250g, nos tocarán 100 presas en la muestra con un promedio de, digamos, 50g. Fíjese que no decimos "imposible" sino "raro" o "extraño". Además, si alguien nos preguntara: "¿cuánto es el peso promedio de la población de presas?", le contestaríamos diciendo el valor que hayamos visto en la muestra y a nuestra afirmación deberíamos agregarle alguna advertencia tal como: "más o menos", o ``aproximadamente".

A un valor calculado con los datos de una muestra para jugar el papel de decir, aproximadamente, el valor de un parámetro de la población, lo denominamos **estimador**. Cuando decimos que se trata de un **estimador puntual** queremos decir que para estimar el parámetro estamos usando un valor único. Volviendo al ejemplo de las presas de los elefantes marinos: si la muestra de 100 presas arroja un valor del promedio de 235 g, diríamos que **estimamos** el promedio de la población en 235 g.

Es decir que dada una población de una variable aleatoria claramente identificada, el proceso de toma de muestras desemboca en el análisis de los valores de dicha variable aleatoria en la muestra con el fin de extraer de ella alguna conclusión acerca de la información contenida en la población, que seguirá siendo objetivamente desconocida. En clases anteriores habíamos definido a las cantidades

calculadas a partir de los datos de la muestra como **estadísticos** y a las cantidades desconocidas contenidas en la población como **parámetros**. Entonces, si se habrá de decidir acerca de un parámetro basándose en lo que el estadístico dice, se pueden hacer dos cosas: (i) especular acerca del valor del *parámetro poblacional desconocido* basándose en la información que brinda un *estadístico muestral conocido* o, (ii) decidir si se acepta que el valor del parámetro es igual, mayor o menor que una cantidad dada. En ambos casos se estará haciendo una **inferencia estadística**. En el primer caso, se estará haciendo una **estimación** del parámetro y al estadístico que se utiliza para estimar al parámetro se le llama, justamente, **estimador**. En el segundo, se estará poniendo a **prueba** una **hipótesis**. En este capítulo nos concentraremos en la estimación de parámetros y en el siguiente trataremos el tema de las pruebas de hipótesis acerca de los parámetros poblacionales.

La estimación de un parámetro puede consistir simplemente en proponer un valor posible para el parámetro basándose en el valor que tiene el estimador, como hicimos en el ejemplo de los pesos de las presas de los elefantes marinos. Este tipo de estimación se denomina **estimación puntual**. Otra manera de estimar un parámetro consiste en proponer, con un grado calculado de *riesgo* de cometer un *error*, un **intervalo** de valores posibles para el parámetro, lo que se denomina **estimación por intervalo**.

Estimación puntual

En lo sucesivo emplearemos el símbolo θ para designar a un parámetro genéricamente, al símbolo $\hat{\theta}$ para designar a su estimador y n será el tamaño de la muestra.

La función matemática que define al estimador será, en general, la misma que define al parámetro. Por ejemplo, si el parámetro desconocido es la proporción (π) de alguna característica en una población de tamaño N - es decir que $\pi = (X/N)$, donde X es la cantidad de unidades que poseen dicha característica en la población – entonces, el estimador será el valor $p = (x/n)$, donde x es la cantidad de unidades que poseen dicha característica en una muestra de tamaño n y p es la proporción de las mismas.

Cuando se tiene una fórmula para estimar y se aplica a una muestra aleatoria, el resultado es aleatorio, es decir los estimadores son variables aleatorias. Como cualquier variable aleatoria, el estimador tiene

- distribución de probabilidad.
- valor esperado: $E(\hat{\theta})$.
- Variancia y desvío standard.

Características deseables en un buen estimador

Ausencia de sesgo en la estimación

Una propiedad muy deseable de un estimador es que su valor esperado coincida con el del parámetro que se pretende estimar. Al menos, quisiéramos que el valor esperado no difiera mucho del parámetro estimado. Por esa razón es importante la cantidad que, técnicamente llamamos **sesgo**. El sesgo es la diferencia entre el valor esperado del estimador y el parámetro que estima:

$$\text{Sesgo} = E(\hat{\theta}) - \theta. \qquad (6.1)$$

Si el sesgo es cero, se dice que el estimador es *insesgado* y ésta es una característica buena para un estimador.

Variancia mínima

Supongamos que $\hat{\theta}_1$ y $\hat{\theta}_2$ son dos estimadores insesgados de θ. Aunque la distribución de cada uno de los dos estimadores tiene media igual a θ las dispersiones de sus valores alrededor de θ podrían ser diferentes. Entre todos los estimadores insesgados de θ, conviene seleccionar aquél que tenga menor variancia.

El $\hat{\theta}$ resultante se denomina **estimador insesgado con variancia mínima** de θ. Así que el estimador insesgado con variancia mínima es el que, entre todos los estimadores insesgados, tendrá mayor probabilidad de producir una estimación cercana al verdadero valor θ.

Estimación consistente

Una vez obtenido un valor para $\hat{\theta}$ a partir de la muestra, es posible que exista una diferencia entre ese valor y el verdadero valor del parámetro (θ). A la diferencia $\hat{\theta}$ - θ se la denomina **error muestral**, y se debe, como su nombre lo indica, a que cuando se toman varias muestras, éstas pueden diferir entre si. Entonces, otra característica deseable en un buen estimador es que las estimaciones que genere estén típicamente cercanas al valor del parámetro, o sea, que tengan baja probabilidad de tener un error muestral importante. Se dice que un estimador es **consistente** si $P\left(\left|\hat{\theta}-\theta\right|>\varepsilon\right)\rightarrow 0$ cuando $n \rightarrow \infty$. En palabras, un estimador es **consistente** si, a medida que aumenta el tamaño de la muestra, la probabilidad del error muestral tiende a ser más pequeña que cualquier cantidad pequeña (ε) que podamos imaginar. Un estimador consistente entonces tiene una alta probabilidad de tomar un valor cercano al valor del parámetro

Métodos de estimación puntual

Hay varios métodos de estimación pero en este curso presentaremos solamente tres. En esta clase veremos dos de ellos (el método de **máxima verosimilitud** y el método de los **momentos**) y en la clase correspondiente a Regresión Lineal veremos el otro (el método de **mínimos cuadrados**).

El método de los momentos

Se denomina **momento de orden 1** de una distribución de probabilidades para una variable discreta X, o de una función de densidad para una variable continua X, al valor de $E(X)$. Análogamente, el **momento de orden 2** de tales funciones será $E(X^2)$. Los momentos pueden estar centrados en algún número de la distribución. Por ejemplo, el **momento de orden 2 centrado en la media** es $E[(X - \mu)^2]$ o sea, la **variancia**.

Este es el método más sencillo y directo y consiste, simplemente, en igualar los momentos de orden 1 y 2 muestrales a los correspondientes momentos poblacionales y, de allí, despejar μ y σ.

Ejemplo. Se efectúan 100 lanzamientos de 3 monedas y se obtienen los siguientes resultados: 11 veces resultó en 3 cruces, 36 veces resultó en 2 cruces y 1 sol, 38 veces resultó en 1 cruz y 2 soles y 15 veces resultó en 3 soles. Obtenga la estimación del parámetro π de la correspondiente distribución binomial de la variable *número de soles*.

Calculamos la media muestral de la variable y la igualamos a la media poblacional:

$$\bar{x} = \frac{0 \cdot 11 + 1 \cdot 36 + 2 \cdot 38 + 3 \cdot 15}{100} = 1.57;$$
$$\begin{aligned} \hat{\mu} &= n \cdot \pi \\ &= 3 \cdot \pi \\ &= 1.57 \Rightarrow \hat{\pi} = 0.52\overline{3} \end{aligned} \quad (6.2)$$

El método de máxima verosimilitud

Lo que caracteriza al método de MV es que provee estimadores consistentes aunque no siempre proporciona estimadores insesgados. Lo presentaremos mediante un ejemplo.

Supóngase que se obtiene una muestra de 10 plantas de las cuales la segunda, la tercera y la octava han florecido mientras que las 7 restantes no lo han hecho. Si designamos a las variables aleatorias que representan a la presencia de flor y a su ausencia como X_i, siendo su valor igual a 1 si hay flor y 0 si no la hay, entonces los valores de las x_i observados en la muestra obtenida son: 0,1,1,0,0,0,0,1,0,0. Por tanto, si la probabilidad de que haya flores es igual a p y la de que no haya flores es igual a $q = 1 - p$, entonces la probabilidad de la muestra observada es igual a: $q \cdot p \cdot p \cdot q \cdot q \cdot q \cdot q \cdot p \cdot q \cdot q = p^3 \cdot q^7$.

La pregunta que nos hacemos al emplear el método de MV es, ¿para que valor de π sería más probable que hubiera ocurrido la muestra que se observó?, es decir, ¿cual es el valor de π que hace que la probabilidad de que ocurra lo que se observó sea máxima? Entonces tenemos que encontrar el valor de π que haga máxima la probabilidad $p^3 \cdot q^7$. Esto se puede hacer tomando logaritmos y derivando con respecto a p:

$$L = \ln(p^3 \cdot q^7) = 3 \cdot \ln(p) + 7 \cdot \ln(q); \quad (6.3)$$

$$\begin{aligned} \frac{dL}{dp} &= \frac{3}{p} - \frac{7}{1-p} \\ &= 0 \Rightarrow p = \frac{3}{10} \Rightarrow \hat{\pi} = \frac{3}{10} \end{aligned}$$

Este es el concepto de **máxima verosimilitud**. Presentaremos directamente los estimadores de MV de los parámetros más comunes.

(a) **Estimador de MV de π.** El estimador de MV de π es la proporción muestral p: $\hat{\pi} = p$ con:

$E(p) = \pi$ y $\sigma_p = \sqrt{\dfrac{\pi \cdot (1-\pi)}{n}}$, así que p es un estimador insesgado y consistente de π.

(b) **Estimador de MV de μ.** El estimador de MV de μ es la media muestral $\bar{x}$: $\hat{\mu} = \dfrac{\sum x_i}{n} = \bar{x}$. La media muestral es un estimador insesgado, consistente y de mínima variancia de la media poblacional.

(c) **Estimador de MV de σ^2.** El estimador de MV de σ^2 de una distribución normal es la variancia muestral, s_{n-1}^2 : $\hat{\sigma}^2 = s_{n-1}^2$. s_{n-1}^2 es un estimador insesgado de σ^2.

Estimación por intervalo

Los estimadores puntuales, con todo lo buenos que pueden ser, no nos proporcionan un valor para el error muestral que se podría estar cometiendo, es decir, sólo obtenemos un valor puntual y ninguna medida del error. En cambio, una estimación por **intervalo de confianza** (de allí su nombre), más que proporcionar un valor puntual, permite obtener un rango o intervalo de valores de los cuales se espera, con un dado margen de confianza, que lleguen a cubrir el verdadero valor del parámetro.

La estimación por intervalo de confianza consiste en la obtención de dos valores extremos, denominados **límite superior** y **límite inferior** del intervalo, que son variables aleatorias. Para establecer dichos límites, se utilizan los datos de una muestra de tamaño *n*. Luego, se establece la probabilidad deseada de que dicho intervalo alcance a cubrir el verdadero valor del parámetro (desconocido), lo que se denomina **nivel de confianza** del intervalo y se simboliza 1 - α. Lo que la muestra debe proporcionar es, en primer lugar, la estimación puntual del parámetro ($\hat{\theta}$); luego, se necesita conocer el tamaño de la muestra (*n*) y el desvío standard del estimador. En símbolos:

$$P\{\hat{\theta} - h \cdot \sigma(\hat{\theta}) \leq \theta \leq \hat{\theta} + h \cdot \sigma(\hat{\theta})\} = 1 - \alpha \qquad (6.4)$$

Como puede verse en la expresión, el intervalo de confianza es simétrico con límite inferior igual a $\hat{\theta} - h \cdot \sigma(\hat{\theta})$ y límite superior igual a $\hat{\theta} + h \cdot \sigma(\hat{\theta})$, ambos variables aleatorias. El factor de confianza *h* es una cola de la distribución por muestreo del estimador: puede ser una distribución normal, una *t* de Student, etc.

Como puede deducirse de la expresión general, el ancho del intervalo de confianza, o sea el valor de $h \cdot \sigma(\hat{\theta})$ depende de *h* y de $\sigma(\hat{\theta})$, el desvío standard del estadístico muestral el cual, a su vez, dependerá de manera inversamente proporcional del tamaño muestral *n*. Es decir que cuanto mayor sea el tamaño de muestra, menor será el ancho del intervalo de confianza (el intervalo de confianza será más preciso) y, a su vez, cuanta mayor sea la confianza que se desea tener (o sea, cuanto menor α se emplee) mayor será el ancho del intervalo.

Intervalo de confianza para la media poblacional

Caso 1: variancia poblacional conocida y variable aleatoria con distribución normal

Hemos anticipado ya que la media muestral, $\bar{x}$, es un estimador puntual insesgado, consistente y de mínima variancia de la media poblacional, μ. También vimos, en una clase anterior, que el desvío standard de este estimador es $\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}$. Si la distribución de la población es normal, o la muestra es grande, de manera que se aplique el Teorema Central del Límite, el intervalo con una confianza 1 - α, será:

$$P\left\{\bar{x} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha \qquad (6.5)$$

Ejemplo.
Una muestra aleatoria de 50 calificaciones en Matemática mostró una media de 75. Se sabe que el desvío estándar poblacional es igual a 10.

(a) Construir un intervalo de confianza del 95% (IC_{95}) para la media poblacional.
(b) ¿Con qué grado de confianza se puede decir que la media de las notas es 75 ± 1?

(a) Dado que se conoce el desvío estándar de la población, usamos la distribución normal:

$$P\left\{\bar{x} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

$$75 - z_{1-\alpha/2} \cdot \frac{10}{\sqrt{50}} \leq \mu \leq 75 + z_{1-\alpha/2} \cdot \frac{10}{\sqrt{50}}$$

Dado que el nivel de confianza es del 95% resulta que $1 - \frac{\alpha}{2} = 0.975$, así que $z = 1.96$ y:
$75 - 1.96 \cdot \frac{10}{\sqrt{50}} \leq \mu \leq 75 + 1.96 \cdot \frac{10}{\sqrt{50}}$ o sea que $72.23 < \mu < 77.77$ es el IC_{95} buscado.

(b) Aquí hay que averiguar el valor de z tal que se obtenga un valor de $z_{1-\alpha/2} \cdot \frac{10}{\sqrt{50}}$ igual a 1. Luego:

$$z_{1-\alpha/2} = 0.707$$
$$\Rightarrow \frac{\alpha}{2} = 0.24$$
$$\Rightarrow \alpha = 0.48$$
$$\Rightarrow 1 - \alpha = 0.52.$$

Caso 2: varianza poblacional desconocida y variable aleatoria con distribución normal

Si el desvío estándar de la población es desconocido se usa al desvío estándar de la muestra, s_{n-1}, como estimador del desvío standard poblacional σ. En este caso, si la distribución de la variable aleatoria es normal, o la muestra es grande, de manera que se aplique el Teorema Central del Límite, en lugar de utilizar z como estadístico en el intervalo, utilizamos la distribución t de Student, con n - 1 grados de libertad:

$$P\left\{\bar{x} - t_{1-\alpha/2;n-1} \cdot \frac{s_{n-1}}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2;n-1} \cdot \frac{s_{n-1}}{\sqrt{n}}\right\} = 1 - \alpha \,. \qquad (6.6)$$

Esta situación, en la cual la verdadera no se conoce y sólo se cuenta con una estimación de ella es la habitual en la mayor parte de las aplicaciones relacionadas con ciencias agropecuarias y ambientales.

Ejemplo.
Supongamos que deseamos obtener una estimación por intervalo de la longitud promedio de cariopse en una variedad de maíz colorado. Podemos extraer primero una muestra aleatoria de, por ejemplo, 17 cariopses para

observación. Supongamos, además que encontramos, que $\bar{x}$ = 10 mm y que s_{n-1} = 0.3 mm. Con estos datos, puede construirse el siguiente IC_{95}:

$$\bar{x} \pm t_{16;0.975} \cdot \frac{s_{n-1}}{\sqrt{n}}$$

$$= 10 \pm 2.120 \cdot \left(\frac{0.3}{\sqrt{17}} \right)$$

o sea $9.846 \leq \mu \leq 10.154$.

Intervalo de confianza aproximado para una proporción poblacional

La distribución normal puede ser usada también para calcular intervalos de confianza aproximados para proporciones, es decir, para estimar la probabilidad de éxito de un experimento binomial. Dada una muestra de tamaño n extraída de una población con distribución binomial cuyo parámetro es π, el estimador puntual de π es la proporción de éxitos observada en las n observaciones. Llamaremos p a este estimador

$$\hat{\pi} = p = \frac{x}{n}$$

El estimador puntual de la varianza de p, que surge de la varianza de una variable binomial, es

$$Var(p) = \frac{p(1-p)}{n}$$

Con estos dos estimadores puntuales y haciendo el supuesto de que la distribución de probabilidad del estimador p puede ser aproximada con una distribución normal, podemos construir el siguiente intervalo de confianza para π:

$$p - z_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} \leq \pi \leq p + z_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}.$$

Esta aproximación es buena cuando el supuesto de que la distribución de probabilidad del estimador p se parece mucho a una distribución normal es razonable. Para que esto sea así, n debe ser grande (p.ej. >30)

Ejemplo.
Una encuesta hecha a una muestra aleatoria de 100 electores mostró que el 59% de ellos está a favor de un candidato. Hallar el IC_{95} para la proporción de todos los electores que están a favor de dicho candidato.

$$\sqrt{\frac{p \cdot (1-p)}{n}}$$

Aquí p = 0.59 y $= \sqrt{\frac{0.59 \cdot (1-0.59)}{100}}$.

$$= 0.0492$$

Luego:

$$0.59 - 1.96 \cdot 0.0492 \leq \pi \leq 0.59 + 1.96 \cdot 0.0492,$$

o sea que $0.494 \leq \pi \leq 0.686$.

Determinación del tamaño de muestra (n) para un grado de precisión deseado

La amplitud de un intervalo de confianza para la media se relaciona con el **grado de precisión** de la estimación. Cuando la varianza es conocida, este grado de precisión está dado entonces por la expresión:

$$e = z_{,1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

A partir de esta expresión podemos calcular qué tamaño de muestra n es necesario para obtener el nivel de precisión deseado e con un nivel de confianza α dado.

$$n = \frac{z_{1-\alpha/2}^2 \cdot \sigma^2}{e^2} \qquad (6.7)$$

Cuando no se conoce la varianza se puede hacer un cálculo aproximado reemplazando σ por su estimador s y el valor de z percentil correspondiente de la distribución t de Student.

Intervalo de confianza para una diferencia entre dos medias con muestras independientes y varianzas poblacionales desconocidas pero supuestamente iguales

A veces, como se dijo antes, el interés central no está en la estimación de un promedio (μ) sino en la estimación de una diferencia entre promedios ($\Delta\mu$). Similarmente al caso del IC para una media poblacional con varianza poblacional desconocida, la diferencia entre dos medias se distribuye como una t de Student. En este caso, el número de grados de libertad es igual a $n_1 + n_2 - 2$.

$$\Delta\bar{x} - t_{n_1+n_2-2;1-\alpha/2} \cdot s_a \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq \Delta\mu \leq \Delta\bar{x} + t_{n_1+n_2-2;1-\alpha/2} \cdot s_a \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \qquad (6.7)$$

donde

$$s_a = \sqrt{\frac{(n_1-1)\cdot s_{(n-1),1}^2 + (n_2-1)\cdot s_{(n-1),2}^2}{n_1+n_2-2}} \qquad (6.8)$$

es el desvío standard amalgamado entre los desvíos standard de las dos muestras.

Ejemplo.
Nos interesan las diferencias entre los rendimientos promedios de maíz (en Kg/Ha) de dos localidades, A y B. A partir de una muestra aleatoria de 12 establecimientos de la localidad A (n_1) y 15 establecimientos de la localidad B (n_2) obtenemos los siguientes estimadores de la media y de la varianza::

$$\bar{x}_1 = 6000,\ s_{(n-1),1}^2 = 565000, \bar{x}_2 = 5400 \text{ y } s_{(n-1),2}^2 = 362500,$$

así que $\Delta\bar{x} = 6000 - 5400 = 600$; y:

$$s_a = \sqrt{\frac{11 \cdot 565000 + 14 \cdot 362500}{12 + 15 - 2}}$$
$$\cong 672.012$$

Luego, el IC_{95} es:

$$\Delta\bar{x} - t_{n_1+n_2-2;1-\alpha/2} \cdot s_a \cdot \sqrt{\frac{n_1+n_2}{n_1 \cdot n_2}} \leq \Delta\mu \leq \Delta\bar{x} + t_{n_1+n_2-2;1-\alpha/2} \cdot s_a \cdot \sqrt{\frac{n_1+n_2}{n_1 \cdot n_2}}$$

o sea

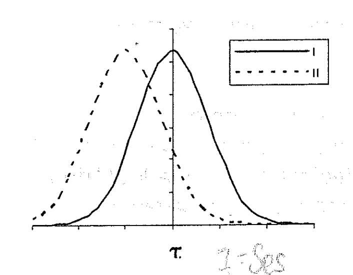
$$63.86 \leq \Delta\mu \leq 1136.14.$$

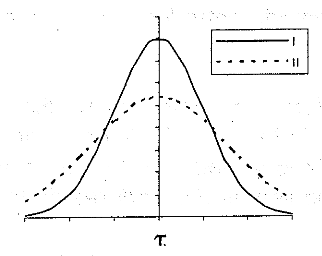
Ejercicios

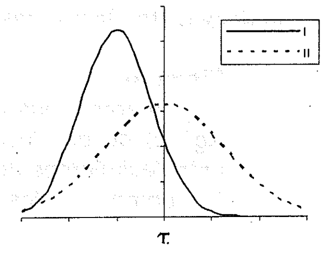
6.1 Decimos que la media muestral es un estadístico porque es una función de los valores de una variable aleatoria medidos en las diferentes unidades muestrales que integran una muestra aleatoria. Como tal, la media muestral es también una variable aleatoria. Para muestras grandes, su distribución de probabilidad depende fundamentalmente del tamaño de la muestra y de la media y la varianza de la variable aleatoria medida. Las características de la distribución de probabilidad de la media muestral hacen que este estadístico sea un sea un estimador insesgado y consistente de la media poblacional de la variable medida.

a. *¿Qué distribución de probabilidad aproximada tiene la media muestral obtenida a partir de una muestra grande?*
b. *¿Qué significa estimador insesgado?*
c. *¿Qué significa estimador consistente?*

6.2 En cada uno de los siguientes diagramas, los números I y II representan las distribuciones muestrales de dos estadísticos que pueden usarse para estimar al parámetro τ. En cada caso, identifique el estadístico que considere como el mejor estimador y justifique su elección.







6.3 La producción ganadera es un problema para la conservación de la fauna natural de los ojos de agua (pequeñas lagunas) de la región húmeda del oeste de Chubut. Las deyecciones de las ovejas enriquecen el agua en nutrientes y esto a su vez causa la proliferación de algas y produce serias consecuencias para los peces y anfibios de las lagunas. Este proceso es denominado "eutroficación" y una medida de su gravedad es la concentración de clorofila

en el agua. En un estudio sobre este problema, se midió la concentración de clorofila en el agua de 10 lagunas tomadas al azar en los establecimientos de cría ovina del oeste de Chubut. Los datos obtenidos son los siguientes:

Concentración de Clorofila en el agua (microgramos/litro)

342	388	348	296	371	304	368	301	392	331

a. *Identificar a la población, las unidades muestrales, la muestra, y la variable aleatoria consideradas en este problema.*

b. *Calcular un estimador puntual insesgado de la concentración de clorofila esperada en una laguna tomada al azar en un establecimiento de cria ovina del oeste de Chubut.*

c. *¿Se puede decir que el valor calculado corresponde a un estimador insesgado de la concentración de clorofila esperada en una laguna tomada al azar en el este de Chubut? Explicar.*

d. *Calcular un estimador puntual insesgado de la varianza de concentración de clorofila de las lagunas de los establecimientos de cría ovina del oeste de Chubut*

e. *Construir un intervalo del 95% de confianza para la concentración de clorofila esperada en las lagunas de los establecimientos de cría ovina del oeste de Chubut.*

f. *Explicar en una frase qué significa el intervalo de confianza construido.*

g. *Determinar un tamaño de muestra suficiente como para estimar la concentración promedio de clorofila de las lagunas en cuestión con un nivel de confianza de 0,95 y una precisión de al menos 10 microgramos/litro.*

6.4 En los cálculos de un intervalo de confianza, la precisión está relacionada con el valor absoluto de la diferencia entre la media muestral y el límite superior o el límite inferior. Respecto de la situación planteada en este ejercicio indicar como se modificaría la precisión en cada uno de los siguientes casos:

a. *Si el intervalo fuese del 99%.*

b. *Si el tamaño de muestra fuera mayor.*

c. *Si el intervalo de confianza se calculara con otra muestra que, por error, incluyera algunas lagunas ubicadas en establecimientos sin ovejas.*

6.5 Una serie de 10 pruebas de cultivo de un nuevo híbrido de maíz realizadas en sitios elegidos al azar en la Pampa Ondulada produce las siguientes estadísticas: $\bar{x} = 9950\ kg/ha$, $s_{n-1} = 920\ kg/ha$

a. *Construir un intervalo del 95% de confianza para el rendimiento esperado.*

b. *Identificar a la población, la muestra, la variable aleatoria consideradas en este problema.*

c. *¿Qué parámetros han sido estimados puntualmente en este caso?* σ y μ

d. *Explicar en una frase qué significa el intervalo de confianza construido.*

6.6 Para estimar la cantidad de forraje presente en una pastura de 10 has se distribuyeron en ella 25 marcos de 1 m^2 ubicados al azar. Todo el forraje presente dentro de cada uno de los marcos fue cortado, secado y pesado. Con los datos obtenidos, se calculó la media aritmética (412 g) y el estimador del desvío standard (s= 96 g) de los pesos.

a. *Identificar a la población, las unidades muestrales, la muestra, y la variable aleatoria consideradas en este caso*
b. *¿Cómo se interpreta el desvío standard observado? ¿Que causas podría tener?*
c. *Construir un intervalo del 95% de confianza para el peso total de forraje (en tn) presente en la pastura.*
d. *Explicar en una frase qué significa el intervalo de confianza construido.*

6.7 La siguiente planilla muestra las alturas (en centímetros) de una población de 100 personas. La variable sigue una distribución aproximadamente normal.

caso	altura	*caso*	altura	*caso*	altura	*caso*	altura	*caso*	altura
1	186	*21*	168	*41*	140	*61*	176	*81*	165
2	177	*22*	146	*42*	179	*62*	179	*82*	179
3	197	*23*	171	*43*	173	*63*	171	*83*	171
4	183	*24*	171	*44*	164	*64*	179	*84*	176
5	178	*25*	181	*45*	173	*65*	170	*85*	178
6	175	*26*	177	*46*	153	*66*	169	*86*	164
7	163	*27*	183	*47*	167	*67*	167	*87*	188
8	165	*28*	177	*48*	160	*68*	172	*88*	170
9	176	*29*	184	*49*	174	*69*	170	*89*	145
10	181	*30*	167	*50*	161	*70*	175	*90*	176
11	166	*31*	170	*51*	173	*71*	170	*91*	173
12	149	*32*	167	*52*	158	*72*	153	*92*	153
13	175	*33*	178	*53*	173	*73*	152	*93*	164
14	190	*34*	171	*54*	169	*74*	178	*94*	153
15	161	*35*	167	*55*	168	*75*	165	*95*	163
16	181	*36*	158	*56*	163	*76*	173	*96*	169
17	168	*37*	184	*57*	174	*77*	161	*97*	160
18	164	*38*	169	*58*	171	*78*	162	*98*	172
19	164	*39*	168	*59*	189	*79*	178	*99*	172
20	163	*40*	180	*60*	146	*80*	171	*100*	166

a. *Calcular la altura media de todas las personas de esta población (es decir, la altura esperada de una persona de esta población tomada al azar).*
b. *Tomar una muestra al azar de tamaño n = 3 y construir un intervalo del 90% de confianza. ¿El intervalo construido incluye a la media poblacional? Repetir el proceso 10 veces.*
c. *Repetir lo hecho en el punto anterior, con una muestra de tamaño n=6.*

6.8 En cada uno de los gráficos que se presentan a continuación, están representados 10 intervalos de confianza para el rendimiento esperado del cultivo de maíz calculados a partir de muestras de tamaño n = 9 obtenidas de una población con media poblacional μ = 7,0 tn/ha y un desvío típico poblacional σ = 1,0 tn/ha. En uno de los gráficos, los intervalos representados corresponden al 90% y en el otro al 99% de confianza.

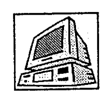
Gráfico A

5,5 6,0 6,5 7,0 7,5 8,0 (tn/ha)

Gráfico B

5,5 6,0 6,5 7,0 7,5 8,0 (tn/ha)

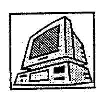
a. *¿Cuál de los gráficos corresponde al 90 y cuál al 99 % de confianza? Justificar la respuesta.*
b. *Explicar por qué los intervalos contenidos en un mismo gráfico son diferentes entre sí.*



6.9 La propaganda de una marca de cigarrillos sostiene que el contenido promedio de nicotina de su producto es menor de 0.7 miligramos por cigarrillo. Para determinar el parámetro toman una muestra al azar de 30 cigarrillos y miden el contenido de nicotina de cada uno de ellos. Los datos obtenidos son los siguientes (en mg/cigarrillo).

0,71	0,75	0,67	0,68	0,72	0,58	0,69	0,70	0,61	0,67	0,74	0,72	0,61	0,63	0,75
0,73	0,59	0,60	0,63	0,59	0,68	0,69	0,77	0,80	0,63	0,62	0,64	0,78	0,76	0,75

a. *Estimar μ con un IC_{99} (intervalo del 99% de confianza)*
b. *Discutir la afirmación que plantea la propaganda sobre la base del intervalo calculado.*



6.10 En una región agrícola se siembra predominantemente una variedad de trigo que tiene un rendimiento medio de 3.5 toneladas por hectárea. Una compañía productora de semillas ha desarrollado una nueva variedad y sostiene que el rendimiento promedio es mayor que en la variedad comúnmente usada. Para evaluar esta aseveración se seleccionan al azar nueve lotes de cultivo dentro de la región y se siembran con la nueva variedad. Los rendimientos que se obtienen figuran en la tabla (en Ton/Ha):

3,15	3,92	4,26	3,72	4,19	3,42	4,38	4,50	3,36

Debemos hacer la suposición que la muestra tiene distribución normal.

a. *Identificar las unidades muestrales, la muestra y la población involucradas en esta prueba.*
b. *Construir un IC_{95} (intervalo del 95 % de confianza).*
c. *Explicar que significa el intervalo construido.*
d. *¿Qué puede decir acerca de la aseveración de la compañía?*

PRUEBAS DE HIPÓTESIS ESTADÍSTICAS

En el capítulo anterior hemos presentado una de las técnicas apropiadas para hacer conjeturas acerca del valor de un parámetro desconocido, la estimación del valor del parámetro. En esta clase, nos referiremos a la segunda técnica que se puede aplicar al decidir si el valor del parámetro es igual, mayor o menor que una cantidad dada: la **prueba de hipótesis**. Básicamente diremos que la técnica de la prueba de hipótesis permite al ingeniero tomar una decisión acerca del valor de un parámetro a partir de la información que puede extraer de una muestra. Esa decisión consistirá en elegir entre dos cursos de acción: dado un valor del estadístico muestral, un valor de dispersión para dicho estadístico y una distribución por muestreo supuesta, se tomará la decisión de rechazar o no que el valor del parámetro pertenece a un conjunto de valores posibles.

Hay dos tipos de hipótesis estadísticas: (i) la **hipótesis nula**, denotada H_0, y, (ii) la **hipótesis alternativa**, denotada H_1. Frente a una situación de incertidumbre acerca del valor de un parámetro (θ), se comienza por plantear una hipótesis que dice que dicho valor (desconocido) corresponde a un dado valor o conjunto de valores (hipótesis nula) y una hipótesis que contempla todos los otros valores posibles. Posteriormente, a través de cálculos basados en la distribución por muestreo del estadístico, se toma o no la decisión de rechazar H_0 – es decir, rechazar que θ es igual al valor que especifica H_0, o que pertenece al conjunto de valores que especifica H_0 y aceptar H_1. Notar que H_0 es rechazada o no es rechazada pero **nunca es aceptada**.

Usualmente, las hipótesis nula y alternativa se plantean en dos formas, según el problema de que se trate:

(a) hipótesis a **dos colas** o **bilateral**. H_0: $\theta = \theta_0$ vs. H_1: $\theta \neq \theta_0$, donde θ_0 es un valor dado de θ.

(b) hipótesis a **una cola** o **unilateral**. La hipótesis unilateral, a su vez, puede ser de dos clases:

(b_1) hipótesis **unilateral izquierda** o hipótesis de **cola izquierda**:

$$H_0: \theta \geq \theta_0 \text{ vs. } H_1: \theta < \theta_0.$$

(b_2) hipótesis **unilateral derecha** o hipótesis de **cola derecha**:

$$H_0: \theta \leq \theta_0 \text{ vs. } H_1: \theta > \theta_0.$$

Ejemplo.
Para lanzar al mercado un nuevo híbrido de maíz, una compañía de semillas debe superar la marca de 11200 Kg/Ha de rendimiento promedio. Para decidir si su última creación genética está en condiciones de salir a competir al mercado, serían apropiadas las siguientes hipótesis:

H_0: $\mu \leq 11200$; si H_0 es cierta, no sale el híbrido nuevo;
H_1: $\mu > 11200$, si H_1 es cierta, sale el híbrido nuevo.

Este es un claro ejemplo de hipótesis de cola derecha, donde la hipótesis nula se rechaza para valores altos (a la derecha de la distribución).

Tipos de error que se pueden cometer cuando se pone a prueba una hipótesis

El hecho de que se tome una decisión acerca del valor de θ , no significa necesariamente que se ha tomado una decisión correcta. La decisión de no rechazar H_0 implica dos resultados posibles: si el verdadero valor de θ pertenece al conjunto de valores especificado por H_0, entonces se ha tomado una *decisión*

correcta, pero si el verdadero valor de θ no pertenece al conjunto de valores especificado por H_0 sino al especificado por H_1, entonces se ha cometido un *error*. Similarmente, la decisión rechazar H_0 implica dos resultados posibles: si el verdadero valor de θ pertenece al conjunto de valores especificado por H_0, entonces se ha cometido un *error*, pero si el verdadero valor de θ no pertenece al conjunto de valores especificado por H_0 sino al especificado por H_1, entonces se tomado una *decisión correcta*. El error de rechazar H_0 cuando es cierta se denomina **error de tipo I** (e_I) y su probabilidad se denota usualmente con la letra α y el error de no rechazar H_0 cuando es falsa se denomina **error de tipo II** (e_{II}) y su probabilidad se denota usualmente con la letra β. Podríamos resumir estas cuatro situaciones en el siguiente cuadro:

Cuadro 7.1.

		H_0 verdadera	H_1 verdadera
Decisión	No se rechazó H_0	Decisión correcta	Error de tipo II
	Se rechazó H_0	Error de tipo I	Decisión correcta

La probabilidad de cometer un error de tipo I es denominada usualmente como α. Frente a un dado planteo de hipótesis, se especifica un valor de α determinado, al que se le denomina **nivel de significación de la prueba**, y una vez calculado el valor de la distribución del estadístico muestral que corresponde a dicha probabilidad (α), al que se denomina **valor crítico**, se toma una decisión. Una vez conocido el valor crítico, el conjunto de valores posibles del estadístico de prueba queda dividido en dos subconjuntos: el conjunto de valores para los cuales no se rechazará H_0 (denominado **región de aceptación**) y el conjunto de valores para los cuales se rechazará H_0 (denominado **región de rechazo**).

Protocolo general de la prueba de hipótesis

Se puede resumir el procedimiento a seguir en las pruebas de hipótesis en los siguientes pasos.

1. Planteo de las hipótesis nula y alternativa.
2. Elección de un nivel de significación para la prueba (α).
3. Elección de un estadístico de prueba. La distribución por muestreo del estadístico de prueba se basa en el supuesto de que H_0 es cierta.
4. Determinación del valor crítico de la prueba en base a α, a la distribución por muestreo del estadístico de prueba y al tipo de hipótesis que se han planteado.
5. Cálculo del valor del estadístico de prueba y su error standard para la muestra que se utilizó y comparar dicho valor con el valor crítico.
6. Decisión: se rechaza o no se rechaza H_0.

Prueba unilateral derecha (H_0: $\theta \leq \theta_0$ vs. H_1: $\theta > \theta_0$)

Supongamos que la distribución por muestreo del estadístico de prueba que se ha elegido es la distribución normal standard (z). Entonces, dado un valor de α, el valor crítico de z correspondiente a dicho nivel de significación ($\hat{\theta}_c$) será $\hat{\theta}_c = \theta_0 + z_{1-\alpha} \cdot \sigma(\hat{\theta})$, donde θ_0 es el valor del estadístico muestral, $z_{1-\alpha}$ es el valor de z correspondiente a la probabilidad 1 - α y $\sigma(\hat{\theta})$ es el valor del error standard del estadístico muestral. En la Figura 1 de la página siguiente se representa el caso de una prueba de cola derecha, con distribución normal standard del estadístico muestral y para dos valores de α: 0.05 y 0.01 donde *R* representa la región de rechazo, *A*, la región de aceptación, 1.65 es el valor de z correspondiente a un valor de probabilidad 1 - α = 0.95 (es decir, α = 0.05), 2.33 es el valor de z correspondiente a un valor de probabilidad 1 - α = 0.99, (es decir, α = 0.01). Así que, para una prueba unilateral derecha, la decisión será, si

usamos $\alpha = 0.05$ (o 0.01), rechazar H_0 si el valor del estadístico muestral (en este caso, z) es superior a 1.65 (o a 2.33). Caso contrario, no rechazar H_0.

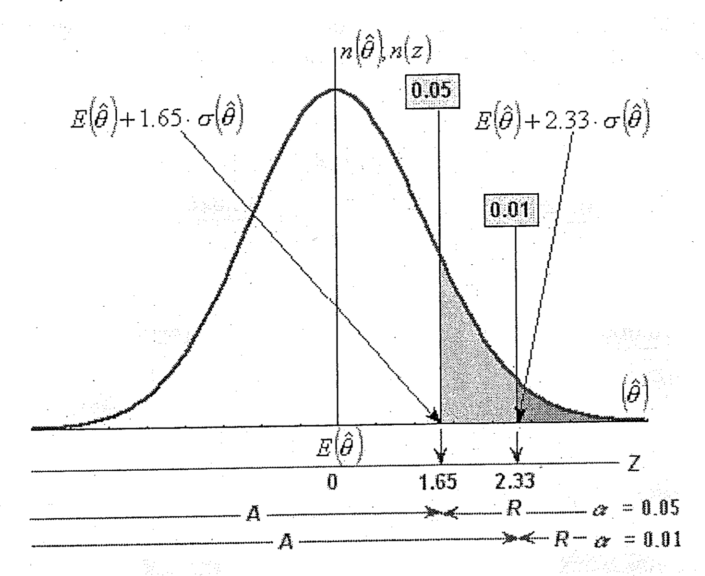


Figura 7.1. Representación esquemática de la región crítica o región de rechazo (R), de la región de aceptación (A) y de las áreas correspondientes a dos niveles de significación, 0.01 y 0.05, para el caso de una prueba unilateral derecha.

Prueba unilateral izquierda (H_0: $\theta \geq \theta_0$ vs. H_1:$\theta < \theta_0$)

Otra vez, supongamos que la distribución por muestreo del estadístico de prueba que se ha elegido es la distribución normal standard (z). Entonces, dado un valor de α, el valor crítico de z correspondiente a dicho nivel de significación ($\hat{\theta}_c$) será

$$\hat{\theta}_c = \theta_0 + z_\alpha \cdot \sigma(\hat{\theta}) \tag{7.1}$$

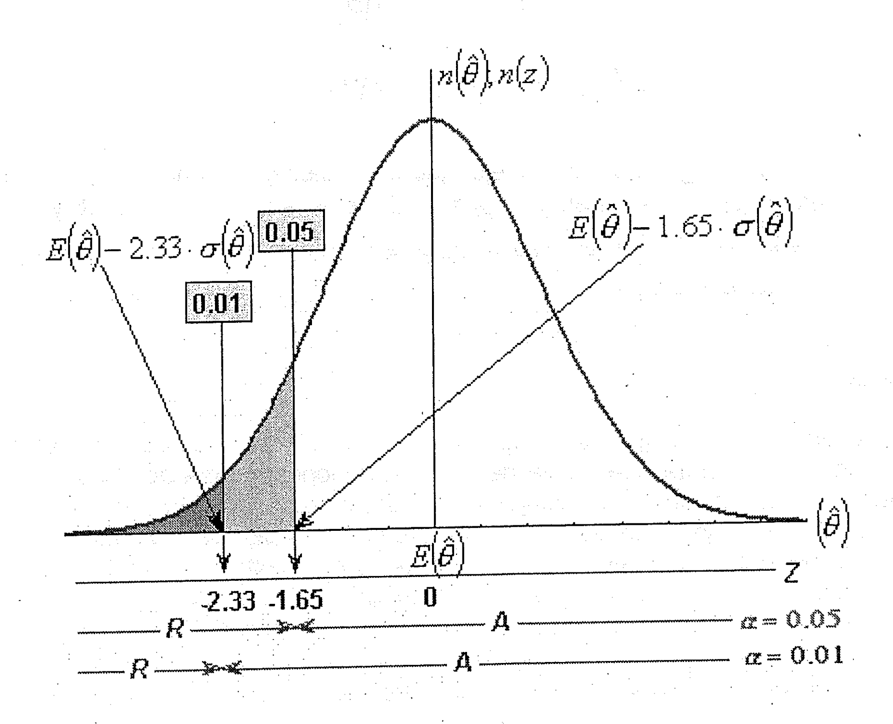


Figura 7.2. Representación esquemática de la región crítica o región de rechazo (R), de la región de aceptación (A) y de las áreas correspondientes a dos niveles de significación, 0.01 y 0.05, para el caso de una prueba unilateral izquierda.

De manera que, para una prueba unilateral izquierda, la decisión será, si usamos $\alpha = 0.05$ (o 0.01), rechazar H_0 si el valor del estadístico muestral (en este caso, z) es inferior a -1.65 (o a -2.33). Caso contrario, no rechazar H_0.

Prueba bilateral (H_0: $\theta = \theta_0$ vs. H_1: $\theta \neq \theta_0$)

En este caso, la región crítica estará dividida en dos segmentos de igual longitud situados (simétricamente) a ambos extremos de la distribución del estadístico (Figura 7.3).

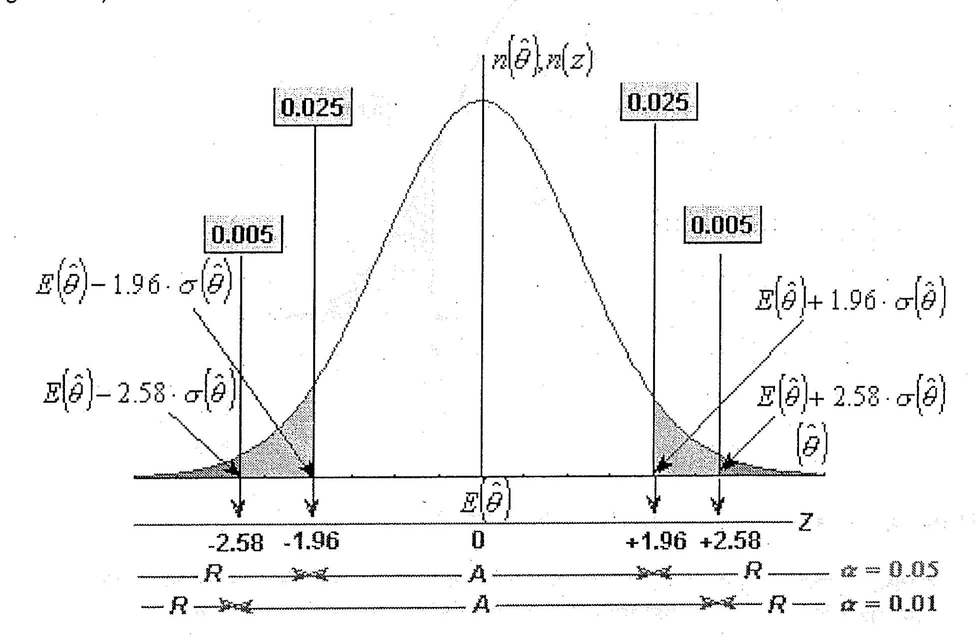


Figura 7.3. Representación esquemática de las dos regiones críticas o de rechazo (*R*), de la región de aceptación (*A*) y de las áreas correspondientes a dos niveles de significación, 0.01 y 0.05, para el caso de una prueba bilateral.

De modo que habrá dos valores críticos, uno a la izquierda y el otro a la derecha:

$$\hat{\theta}_{cI} = \theta_0 + z_{\alpha/2} \cdot \sigma(\hat{\theta}) \quad \text{y} \tag{7.2}$$

$$\hat{\theta}_{cD} = \theta_0 + z_{1-\alpha/2} \cdot \sigma(\hat{\theta}) \tag{7.3}$$

Por ejemplo, siguiendo con el ejemplo de la distribución normal standard, si $\alpha = 0.05$, entonces, $z_{\alpha/2}$ = -1.96 y $z_{1-\alpha/2}$ = +1.96; si α = 0.01, $z_{\alpha/2}$ = -2.58 y $z_{1-\alpha/2}$ = +2.58. Por tanto, la decisión en este caso será no rechazar H_0 si $\hat{\theta}_{cI} < \hat{\theta} < \hat{\theta}_{cD}$. Caso contrario, se rechaza H_0.

El valor *p*

Con el advenimiento del uso de computadoras y de *software* estadístico, se ha generalizado una manera alternativa de tomar decisiones acerca del valor de un parámetro. Frente a un dado conjunto de datos de muestra, el software estadístico calcula el valor del estadístico de prueba y el valor de probabilidad que le corresponde (**valor *p***), según la distribución por muestreo asumida para el mismo. Entonces, en lugar de fijar de antemano un nivel de significación y observar si el valor del estadístico calculado está por debajo o por encima del valor crítico, el ingeniero toma su decisión sobre la base de dicho valor *p*. En este curso, ejemplificaremos el uso de ambas estrategias.

Prueba de hipótesis sobre la media poblacional de una variable con distribución normal

La media poblacional es una medida cuyo conocimiento o, en su defecto, estimación, usualmente, resulta muy necesario. Por ejemplo, un nuevo cultivar de trigo, ¿puede elevar el rendimiento promedio de las cosechas en una determinada localidad, si es adoptado? ¿Se ha elevado el ingreso *per capita* real en la Argentina en el último año? ¿Alcanzó el cultivo de maíz de un lote el límite de humedad necesario para cosecharlo?

En todos estos casos con los datos de una muestra necesitamos extraer conclusiones acerca de la media de la población. Como hemos visto, el estadístico que se emplea para estimar la media poblacional (μ) es la media muestral ($\bar{x}$). Cuando se trata de *una variable con distribución normal o la muestra es suficientemente grande para que opere el teorema central del límite, entonces* el estadístico

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s_{n-1}/\sqrt{n}} \qquad (7.4)$$

t de Student con n-1 grados de libertad siempre y cuando la hipótesis nula $\mu = \mu_0$ sea cierta. Esto permite poner a prueba la hipótesis nula como en el siguiente ejemplo.

Ejemplo.
Supongamos que una máquina enfardadora produce fardos con un ancho de 80 cm. Para controlar el funcionamiento de la máquina se tomó una muestra de 20 fardos en la cual el ancho medio resultó ser de 77 cm con un desvío standard de 12 cm. Probar la hipótesis de que la máquina está trabajando correctamente con $\alpha = 0.10$.

En este caso, se debe considerar que la máquina está trabajando correctamente si produce empaques que no sean demasiado grandes ni demasiado pequeños así que se trata claramente de una prueba bilateral.

1] **Hipótesis**. H_0: $\mu = 80$; H_1: $\mu \neq 80$.
2] **Nivel de significación**. $\alpha = 0.10$.
3] **Estadística de prueba**. $t_{n-1} = \frac{\bar{x} - \mu_0}{s_{n-1}/\sqrt{n}}$ que se distribuye como una t_{19}.
4] **Región crítica.** Puesto que $P(t_{19} < -1.729 \cup t_{19} > +1.729) = 0.10$, se rechazará H_0 si $t < -1.729$ ó $t > +1.729$.
5] **Cálculos.** $n = 20$, $\bar{x} = 77$, $s_{n-1} = 12$ y

$$t_{19} = \frac{77-80}{12/\sqrt{20}} = \frac{-3}{2.683} = -1.118.$$

6] **Decisión**. Dado que el valor del estadístico de prueba no cae en ninguna de las dos regiones críticas, H_0 no es rechazada.

7] Con el menú ***Estadísticas – Probabilidades y cuantiles*** de ***Infostat***, podemos calcular el valor *p* de la prueba. Elegimos $v = 19$, que es el número de grados de libertad. El valor *p* es, aproximadamente, igual a 0.2779 que es muy superior a 0.10.

Pruebas de hipótesis para una proporción poblacional

En los casos en que se analiza una variable con distribución binomial, como por ejemplo el número de plantas enfermas en una muestra aleatoria de tamaño n, frecuentemente interesa poner a prueba hipótesis sobre la proporción de, por ejemplo, plantas enfermas en la población. Denotaremos a esta proporción con la letra π y a su estimador con la letra p. Las hipótesis involucrarán al parámetro π y a algún valor particular de dicho parámetro, π_0.

Prueba exacta basada en la distribución binomial:

Ejemplo.
Se presume que la proporción de plantas atacadas por una enfermedad en un lote experimental, es del 12%. Se toma una muestra de 20 plantas y se halla 1 sola planta enferma en él. Poner la hipótesis de interés utilizando un nivel de significación $\alpha = 0.10$.

1] **Hipótesis**. H_0: $\pi \geq 0.12$; H_1: $\pi < 0.12$ (prueba de cola izquierda).

2] **Nivel de significación**. $\alpha = 0.10$.

3] **Estadística de prueba.** La estadística de la prueba es el *número de plantas enfermas* observadas en la muestra (x). En nuestro caso, $x = 1$.

4] **Región crítica.** La región crítica estará determinada por la cola izquierda de la distribución binomial:

$$p = P(k \leq x) = \sum_{k=2}^{20} \binom{20}{k} \cdot 0.12^k \cdot (0.88)^{20-k}$$

$$= 1 - \sum_{k=0}^{1} \binom{20}{k} \cdot 0.12^k \cdot (0.88)^{20-k}$$

Utilizando el programa ***Infostat***, en el menú ***Estadísticas – Probabilidades y Cuantiles***, se puede elegir la distribución ***binomial*** y, especificando los valores de n y p (en nuestros símbolos no sería p sino π), se pueden obtener todos los valores. En nuestro ejemplo, $n = 20$ y $\pi = 0.12$. En la siguiente tabla se presentan sus valores ya calculados.

X	0	1	2	3	4	5	6	7
P(x)	0.0776	0.2115	0.2740	0.2242	0.1299	0.0567	0.0193	0.0053
X	8	9	10	11	12	13	14	15
P(x)	0.0012	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
X	16	17	18	19	20			
P(x)	0.0000	0.0000	0.0000	0.0000	0.0000			

En este caso, la región crítica, con $\alpha = 0.10$, estará formada por un valor de x entre 0 y 1 puesto que ya para $x = 1$ tenemos ya que $P(x) = 0.2115$. Por tanto, con distribuciones discretas es posible que no se pueda elegir un nivel de significación exacto como se desee. Aquí, el valor más cercano es 0.0776 que corresponde a $x = 0$. Tomaremos $x = 1$ como cota superior.

5]Cálculos

$$p = P(x \leq 1) = \binom{20}{0} \cdot 0.12^0 \cdot (0.88)^{20} + \binom{20}{1} \cdot 0.12^1 \cdot (0.88)^{19}$$

$$= 0.88^{20} + 20 \cdot 0.12 \cdot 0.88^{19}$$
$$= 0.0776 + 0.2115$$
$$= 0.2891.$$

6] Decisión. Dado que el valor de p es bastante mayor a 0.10, no existe evidencia suficiente en la muestra como para rechazar H_0.

7] El valor p de la prueba es justamente la probabilidad calculada en [5]: 0.2891.

Prueba aproximada basada en la distribución normal:

Cuando n, el numero de observaciones realizadas para evaluar un proporción es un número muy grande, puede resultar engorroso calcular probabilidades con la fórmula de la distribución binomial. En ese caso, sin embargo, el estimador p de la proporción poblacional π tiene distribución aproximadamente normal. En ese caso, el estadístico

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \cdot (1 - \pi_0)}{n}}}$$

tiene distribución aproximadamente normal standard, siempre y cuando se cumpla la hipótesis nula $\pi = \pi_0$. Esto provee una alternativa sencilla permite construir una prueba de hipótesis aproximada cuando el número de observaciones n es grande.

Ejemplo.
Supongamos que una fábrica de agroquímicos (X) ha anunciado en un diario lo siguiente: "Hay varios proveedores de agroquímicos en la zona, pero un 42% de los productores usan nuestros productos". Supongamos que los otros fabricantes rechazan esta aseveración. Se debe decidir acerca de la verdad de esta afirmación y, para ello, se toma una muestra, sin reposición, de 240 productores. Se encuentra que 90 de ellos, o sea un 37.5%, usan productos fabricados por X. ¿Es estadísticamente significativa la diferencia de 4.5% entre el resultado de la muestra y la proporción anunciada por el fabricante? (usar $\alpha = 0.01$).

1] Hipótesis. H_0: $\pi \geq 0.42$; H_1: $\pi < 0.42$ (prueba de cola izquierda).

2] Nivel de significación. $\alpha = 0.01$.

3] Estadística de prueba. $z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \cdot (1 - \pi_0)}{n}}}$ que se distribuye como una z.

4] Región crítica. $P(z < -2.327) = 0.01$, así que se rechazará H_0 si $z < -2.327$.

5] Cálculos. $p = \frac{90}{240} = 0.375$,

$$z = \frac{0.375 - 0.42}{\sqrt{\frac{0.42 \cdot (1 - 0.42)}{240}}} = \frac{-0.045}{0.0319} = -1.410.$$

6] Decisión. Dado que el valor de z se encuentra en la región de aceptación, no se rechaza H_0.

7] Con el menú ***Estadísticas – Probabilidades y cuantiles*** de ***Infostat***, podemos calcular el valor p de la prueba. En este caso dicho valor es, aproximadamente, igual a 0.0793 que es superior a $\alpha = 0.010$.

Prueba de hipótesis sobre la diferencia entre las medias de dos variables con distribución normal

Cuando el interés del investigador o del ingeniero no está ya en una media poblacional sino en la diferencia entre dos medias poblacionales, el parámetro poblacional será el parámetro diferencia ($\Delta\mu = \mu_1 - \mu_2$) y su estimador muestral será la diferencia en la muestra $\left(\Delta\bar{x} = \bar{x}_1 - \bar{x}_2\right)$. Según cómo han sido obtenidos los datos, aparecen dos situaciones diferentes para poner a prueba hipótesis acerca del valor de $\Delta\mu$; en la primera situación, las unidades muestrales que integran las dos muestras están apareadas y en el segundo son independientes. Las pruebas de hipótesis apropiadas difieren entre estas dos situaciones.

Muestras apareadas

En algunas situaciones conviene comparar las medias de dos poblaciones a partir de muestras relacionadas de modo tal que las unidades de muestreo formen parejas. Por ejemplo, para comparar el rendimiento medio obtenido con dos híbridos de maíz, cada par estaría constituido por dos lotes de cultivo de una misma localidad y cada miembro del par está cultivado con uno de los híbridos. De esta manera, cada diferencia entre los rendimientos obtenidos en cada localidad constituye un estimador de la diferencia entre los métodos bajo condiciones determinadas existentes en la localidad correpondiente. Los datos que se van a analizar consisten en una muestra de n diferencias los rendimientos en n localidades. El objetivo del muestreo apareado es generar pares que sean lo más homogéneos posible en los factores diferentes del que se está analizando (p.ej, el híbrido de maíz utilizado), de manera de poder atribuir las diferencias encontradas a dicho factor.

En estos casos, la información está formada por n **pares** seleccionados de manera independiente (x_1,y_1), (x_2,y_2), ..., (x_n,y_n), con $E(x_i) = \mu_1$ y con $E(y_i) = \mu_2$. Sea, entonces, la variable $d_i = x_i - y_i$ tal que el valor de d_i sea las diferencias entre ambas muestras dentro del par i. Se partirá del supuesto que las d_i tienen distribución normal con variancia σ_d^2.

Estamos interesados en poner a prueba la hipótesis H_0: $\mu_d = \mu_1 - \mu_2 = \Delta_0$, donde Δ_0 es una diferencia particular. El estadístico a utilizar en la prueba de hipótesis será:

$$t_{n-1} = \frac{\bar{d}}{s_d/\sqrt{n}} \qquad (7.5)$$

que tiene distribución t de Student con $n - 1$ grados de libertad;

$$\bar{d} = \sum_i \frac{d_i}{n}$$

es la media aritmética de las diferencias, donde n es el número de parejas. El estimador del desvío standard de esta media aritmética de las diferencias es

$$\frac{s_d}{\sqrt{n}}$$

con:

$$s_d = \sqrt{\frac{\sum_i (d_i - \bar{d})^2}{(n-1)}}$$

Ejemplo.
Supongamos que se desea saber si un nuevo híbrido de maíz (*B*) es superior a otro híbrido anterior (*A*) por su rendimiento promedio en 10 localidades de la región maicera de la provincia de Buenos Aires. Se eligieron al azar 10 establecimientos y se obtuvieron los resultados que se presentan el cuadro siguiente en el cual ya se han calculado las diferencias para cada localidad y sus respectivos cuadrados.

Cuadro 7.2.

Localidad	Híbrido *A*	Híbrido *B*	d_i	$(d_i)^2$
I	8450	8239	+211	44521
II	7929	8130	-201	40401
III	8126	8255	-129	16641
IV	8847	8750	+97	9409
V	9059	9147	-88	7744
VI	8732	8643	+89	7921
VII	8346	8442	-96	9216
VIII	8009	8112	-103	10609
IX	8859	9047	-188	35344
X	8642	8540	+102	10404
Total	**84999**	**85305**	**-306**	**192210**

1] Hipótesis. H_0: $\mu_1 \geq \mu_2$; H_0: $\mu_1 < \mu_2$.

2] Nivel de significación. $\alpha = 0.01$.

3] Estadística de prueba. $t_{n-1} = \dfrac{\bar{d}}{s_d/\sqrt{n}}$ que tiene distribución *t* de Student con $n - 1$ grados de libertad.

4] Región crítica. Para $n = 10$, obtenemos que $P(t_9 < -2.821) = 0.01$ y se rechazará H_0 si, y solo si, $t_9 < -2.821$.

5] Cálculos. $\bar{d} = \sum_i \dfrac{d_i}{n} = \dfrac{-306}{10} = -30.6$;

1] $s_d = \sqrt{\dfrac{\sum (d_i - \bar{d})^2}{(n-1)}} = \sqrt{\dfrac{10 \cdot 192210 - (-306)^2}{10 \cdot (10-1)}} = 142.535$;

2] $\dfrac{s_d}{\sqrt{n}} = \dfrac{142.535}{\sqrt{10}} = 45.074$; $t_{n-1} = \dfrac{-30.6}{45.074} = -0.679$

6] Decisión. Puesto que -0.679 > -2.281, H_0 no es rechazada y concluimos en que no hay diferencias entre las medias de rendimiento de los dos híbridos de maíz, en esta región.

7] Calculamos el valor *p* de la prueba con ***Infostat*** con $v = 9$ grados de libertad. El valor *p* es, aproximadamente, igual a 0.2571 que es muy superior a $\alpha = 0.010$.

También se puede calcular un ***intervalo de confianza para la media de las diferencias***, por ejemplo podemos calcular un IC_{99} para $\Delta\mu$:

$\bar{d} \pm t_{n-1;1-\alpha/2} \cdot \frac{s_d}{\sqrt{n}}$ = -30.6 ± 3.250·45.074 o sea: -177.09 ≤ $\Delta\mu$ ≤ 115.88.

Este ejercicio puede ser realizado con ***Infostat***. Para ello se deben cargar los datos de rendimiento de los dos híbridos en dos columnas distintas. Luego se debe recurrir al menú **Estadísticas – Inferencia basada en dos muestras – Prueba *t* apareada** y, allí, elegir como **Variables**, a la Columna 1 y a la Columna 2. Luego, tildar en la casilla ***Intervalo de Confianza*** indicando 99 en la casilla para el nivel de confianza.

Muestras independientes

En este punto tratamos con muestras tomadas independientemente una de la otra. Consideraremos sólo el caso en el cual las poblaciones de las cuales provienen las muestras tienen igual variancia. En este caso, nuestro estimador insesgado de dicha variancia (que es la misma para ambas poblaciones) es:

$$s_a^2 = \frac{(n_1 - 1) \cdot s_{(n-1),1}^2 + (n_2 - 1) \cdot s_{(n-1),2}^2}{n_1 + n_2 - 2} \qquad (7.6)$$

Este estimador, frecuentemente denominado *la variancia amalgamada*, es un promedio ponderado (amalgamado) de los estimadores de la variancia derivados de las dos muestras. Consecuentemente, el estimador del desvío standard (o error standard) de la diferencia entre las medias aritméticas muestrales es

$$s_a \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (7.7)$$

En este caso, el estadístico

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta\mu_0}{s_a \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (7.8)$$

tiene distribución *t* de Student con $n_1 + n_2 - 2$ grados de libertad siempre y cuando la hipótesis nula que dice $\Delta\mu = \Delta\mu_0$ sea cierta. Esto permite poner a prueba la hipótesis nula como en el ejemplo que sigue.

Ejemplo.
En una estación experimental agropecuaria se desea evaluar el efecto de cierto herbicida sobre la producción de cebada. Con ese fin, se seleccionan 28 parcelas de tierra, a 14 de ellas se las trata con herbicida y a las otras 14 no. La producción promedio de cebada de las parcelas no tratadas fue de 5 toneladas con un desvío standard igual a 0.5 toneladas. La producción promedio de las parcelas tratadas fue de 5.3 toneladas con un desvío standard igual a 0.7 toneladas. Extraer una conclusión con α = 0.05 y determinar el valor *p* de la prueba de hipótesis.

Cuadro 7.3.

Con herbicida	Sin herbicida
$n_1 = 14$	$n_2 = 14$
$\bar{x}_1 = 5.3$	$\bar{x}_2 = 5.0$
$s_1 = 0.7$	$s_2 = 0.5$

Podemos resumir la información muestral así:

1] Hipótesis. H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$;

2] Nivel de significación. $\alpha = 0.05$.

3] Estadística de prueba. $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s_a \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ que tiene distribución t_v

donde $v = n_1 + n_2 - 2 = 26$.

4] Región crítica. Con $\alpha = 0.05$ para una prueba bilateral: $t_{26} < -2.056$ y $t_{26} > +2.056$. Por tanto, se rechazará H_0 si $t_{26} < -2.056$ o $t_{26} > +2.056$.

5]Cálculos.

$$s_a = \sqrt{\frac{(n_1 - 1) \cdot s^2_{(n-1),1} + (n_2 - 1) \cdot s^2_{(n-1),2}}{n_1 + n_2 - 2}} = \sqrt{\frac{13 \cdot 0.49 + 13 \cdot 0.25}{14 + 14 - 2}} = 0.608$$

$$s_a \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 0.608 \cdot \sqrt{\frac{1}{14} + \frac{1}{14}} = 0.230 \text{ y}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{0.230} = \frac{5.3 - 5.0}{0.230} = 1.304.$$

6] Decisión. El valor de t calculado no es significativo (no cae en la región de rechazo de la hipótesis nula). Al 5% de significación se decide no rechazar la H_0, es decir que no hay evidencias de un efecto del herbicida estadísticamente significativo sobre los rendimientos.

7] Ahora calculamos el valor p de la prueba con ***Infostat***. Elegimos $v = 26$, que es el número de grados de libertad. El valor p es, aproximadamente, igual a 0.2036 que es muy superior a 0.05.

Ejercicios

7.1 En 1996, un establecimiento ganadero de la provincia de Chubut puso en marcha un plan de mejoramiento de la calidad de la lana basado en la incorporación de machos reproductores de una reconocida cabaña. Un censo de todas las ovejas presentes en el establecimiento mostró que, en ese momento, la media poblacional de la longitud de la lana del vellón era de 79,7 mm. El ingeniero responsable del plan considera que si el plan de mejoramiento ha sido efectivo dicha media poblacional debería haber aumentado luego de 10 años. En consecuencia, en 2006, este ingeniero toma del rodeo10 ovejas al azar y les mide la longitud de la lana del vellón. Los datos que obtiene son los siguientes.

Oveja	1	2	3	4	5	6	7	8	9	10
Longitud de lana (mm)	80,9	80,0	80,7	77,3	81,9	78,1	81,8	81,1	79,5	79,0

a. *Realizar una prueba de hipótesis para tomar una decisión respecto de la siguiente afirmación: " si el plan de mejoramiento no fue efectivo y en consecuencia se deberá cambiar de cabaña proveedora de machos reproductores". Presentar el análisis y la conclusión.*

b. *¿Cual sería el impacto para la empresa de cometer un error de tipo I o un error de tipo II?*

7.2 Una empresa productora de semillas ofrece un nuevo híbrido de maíz que a sido puesto a prueba en 12 lotes tomados al azar dentro del partido de Pergamino. Los rendimientos obtenidos en cada lote fueron los siguientes:

Lote	1	2	3	4	5	6	7	8	9	10	11	12
Rendimiento (tn/ha)	7.2	10.0	8.5	8.4	8.0	7.5	9.0	9.0	8.0	7.0	6.1	8.0

a. *Identificar la poblacion, la muestra y la variable aleatoria consideradas*

b. *Constriur un diagrama de caja y bigotes para los datos de rendimiento de los lotes.*

En Pergamino el costo de producción del maíz híbrido es de U$S 325/ha y el ingreso neto por cada tonelada vendida es U$S 50.

c. *¿Puede asegurarse con un nivel de confianza de al menos 99% que el rendimiento esperado alcanza para cubrir el costo de producción? (Mostrar el desarrollo del análisis)*

d. *Sobre la base del resultado obtenido, discutir brevemente la conveniencia de adoptar este nuevo híbrido en el partido de Pergamino.*

7.3 En un establecimiento lechero se proyecta utilizar el pasto presente en una pastura de 20 has para hacer una reserva de fardos de heno con la cual alimentar a las vacas lecheras durante el invierno. Por ello es muy importante determinar si los fardos a producir alcanzarán para cubrir las 90 tn de forraje que serán necesarias durante dicho período. Para hacer dicha determinación toma una muestra de 25 marcos de 1 m^2 elegidos al

azar dentro de la pastura. En cada marco, se corta todo el forraje y lo pesa luego de dejarlo secar al aire del mismo modo que se hace para elaborar los fardos. El promedio de los pesos obtenidos es de 510 gramos/m^2 y el estimador del desvío standard es de 100 g/m^2.

a. *Identificar las unidades muestrales, la muestra y la población.*
b. *Formular hipótesis apropiadas para evaluar si la cantidad de pasto es suficiente.*
c. *Calcular el nivel de confianza en la hipótesis nula (valor p).*
d. *Concluir con un nivel de significación α = 0,05.*
e. *Explicar que conclusión debería extraerse.*
f. *Explicar el tipo de error que se puede haber cometido en este análisis y cuales serían sus implicancias*

7.4 La propaganda de cierta marca de cigarrillos sostiene que el contenido promedio de nicotina de su producto es menor de 0.7 miligramos por cigarrillo. Suponiendo que el contenido de nicotina de un cigarrillo tomado al azar es una variable aleatoria con distribución normal, su aseveración es que $\mu < 0.7$. Entonces, se desea probar: Ho: $\mu \geq 0.7$ en oposición a H1: $\mu <$ 0.7

La hipótesis se quiere probar con un nivel de significación igual a 0.01, ya que si se rechaza Ho se deberá autorizar que en la publicidad aparezca esta afirmación, y solo estamos dispuestos a hacerlo si la evidencia en contra de Ho es fuerte. Para realizar la prueba determinamos el contenido de nicotina en 30 cigarrillos tomados al azar. Los valores encontrados son los siguientes (mg/cigarrillo):

0.71	0.75	0.67	0.68	0.72	0.58	0.69	0.70	0.61	0.67	0.74	0.72	0.61	0.63	0.75
0.73	0.59	0.60	0.63	0.59	0.68	0.69	0.77	0.80	0.63	0.62	0.64	0.78	0.76	0.75

¿Cuál es la conclusión? Compare estos resultados con los obtenidos en el ejercicio 6.9 del capítulo anterior.

7.6 Para evaluar la exactitud de una nueva técnica para medir el contenido de Arsénico en el agua, un químico prepara una solución que contiene exactamente 50.10^{-3} mg de Arsénico /l. Luego toma 9 alícuotas y en cada una mide el contenido de arsénico x con la técnica propuesta y calcula el error de medición $\delta = x\ mg/l - 50.10^{-3}\ mg/l$. Con estos datos calcula la media aritmética $\bar{\delta} = 1{,}18.10^{-3}\ mg/l$ y el estimador del desvío típico $s = 1{,}52\ 10^{-3}\ mg/l$ de δ.

a. *Identificar las unidades muestrales, la muestra y la población.*
b. *¿Puede concluirse, con un nivel de significación α = 0,05, que el valor esperado de δ es mayor que cero?*
c. *Explicar qué es el nivel de significación α = 0,05 en términos de este problema.*

7.7 Un consorcio de productores agrícolas (CREA) lleva adelante un estudio para comparar los rendimientos de maíz obtenidos con dos métodos de

cultivo diferentes, labranza mecánica y labranza química. Para ello, cada socio del CREA toma un lote que ha sido cultivado como una unidad al menos en los últimos 5 años, lo divide en dos y cultiva maíz aplicando uno de los dos tipos de labranza en cada mitad. Al final de la campaña, los productores logran reunir la siguiente información:

Rendimiento de maíz [tn/ha]

Productor	1	2	3	4	5	6	7	8	9	10
Labranza mecánica	8.9	7.8	10.1	9.7	9.2	9.1	9.9	8.4	9.0	7.2
Labranza química	8.8	6.8	12.9	11.9	8.0	12.2	9.1	11.2	10.5	10.1

a. Estimar el promedio y la varianza de las diferencias de rendimiento entre métodos de cultivo

b. Nombrar posibles causas de la varianza en la diferencia de rendimiento entre métodos de cultivo.

c. Fromular hipótesis para evaluar si los dos métodos de cultivo producen en promedio igual rendimiento.

d. Calcular el nivel de confianza en la hipótesis nula (valor p).

e. Concluir con un nivel de significación a = 0,05.

f. Explicar la conclusión en términos del objetivo del estudio propuesto por el CREA.

7.8 El contenido de gluten en el trigo puede ser afectado no sólo por su tratamiento posterior a la cosecha, sino también por la cantidad de nitrógeno que las plantas pueden absorber en diferentes etapas de su desarrollo. Para evaluar la importancia relativa de la disponibilidad de nitrógeno temprano y tarde en el ciclo del cultivo, se tomaron 10 parcelas sembradas con trigo y cada una fue dividida en dos. Una mitad de cada parcela fue fertilizada en el momento de la siembra y la otra mitad fue fertilizada recién cuando las plantas florecieron. Al final del cultivo se determinó el contenido de gluten del trigo cosechado en cada media parcela mediante la medición la elasticidad de la masa producida con la harina correspondiente. Los datos obtenidos figuran en la tabla:

Elasticidad de la masa (Valor W)

Momento de fertilización	*Parcela*									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A la Siembra	168	138	153	152	159	180	147	159	175	150
A la Floración	188	195	147	178	177	178	179	172	177	185

a. Formular las hipótesis necesarias para evaluar, a partir de estos datos, si el contenido esperado de gluten difiere entre trigo fertilizado a la siembra o a la floración.

g. ¿ Calcular el nivel de confianza en la hipótesis nula (valor p).

h. Concluir con un nivel de significación a = 0,05.

b. Explicar la conclusión en términos del problema.

c. Explicar el tipo de error que se puede haber cometido en esta prueba.

d. *Construir un intervalo del 95% de confianza para la diferencia entre los contenidos de gluten de trigo producido con fertilización a la siembra y a la floración*

7.9 Durante la última década, una importante superficie de los pastizales de la Región Pampeana ha sido reemplazada por forestaciones. Este cambio en el uso de la tierra puede producir consecuencias ambientales debidas a modificaciones de la hidrología local, como cambios en el caudal de los arroyos, en el contenido de sales del suelo o en la profundidad de la napa freática (agua subterránea). En un estudio orientado a evaluar el impacto de las forestaciones sobre el ciclo hidrológico, se seleccionaron al azar 10 forestaciones en el partido de Zárate y, en cada una ellas, se midió la profundidad de la napa freática (en metros) en el centro de una forestación y en un pastizal vecino. Los datos obtenidos figuran en la tabla:

	1	2	3	4	5	6	7	8	9	10
Forestación	2,0	2,3	2,2	2,0	2,3	2,5	2,0	2,3	2,4	2,0
Pastizal	1,5	1,6	1,6	1,5	1,8	1,8	1,5	1,6	1,7	1,4

a. *Identificar la población, la muestra y las unidades de observación y las variables aleatorias involucradas en este estudio.*

b. *Estimar la esperanza y la varianza de las diferencias en la profundidad de napa entre pastizales y forestaciones.*

c. *Construir un intervalo del 95% confianza para la esperanza de dichas diferencias*

d. *¿Se puede concluir con un nivel de significación α=0,05 que, en Zárate, las forestaciones han determinado un aumento en la profundidad promedio de la napa freática?*

e. *¿Qué tipo de error se puede haber cometido en la prueba de hipótesis anterior? Explicar su significado en términos del problema.*

7.10 En un estudio sobre la susceptibilidad de plántulas de duraznero a dos cepas diferentes de un virus, se tomaron de un vivero 8 plántulas al azar; en cada plántula se seleccionaron 2 hojas y cada una fue inoculada con una de las dos cepas virales. Al cabo de una semana, se midió en cada hoja el tamaño de la lesión producida por el virus (en mm^2). Los datos obtenidos figuran en la tabla:

Planta	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Lesión cepa viral A [mm^2]	31	20	18	17	9	8	10	24
Lesión cepa viral B [mm^2]	18	17	14	11	10	7	5	25

a. *Estimar el promedio y la varianza de las diferencias entre los tamaños de las lesiones producidas por las dos cepas virales estudiadas.*

b. *Poner a prueba, con un nivel de significación α=0,05, la siguiente hipótesis nula: Las dos cepas virales producen lesiones con el mismo promedio de tamaño.*

c. *Explicar qué es el nivel de significación α.*

7.11 Para estudiar el efecto del cobre sobre la ganancia diaria de peso de terneros, se tomó una muestra aleatoria de 12 terneros de un establecimiento ganadero donde los suelos son deficientes en cobre. A 5 terneros seleccionados al azar se les aplicó a un tratamiento de inyección de cobre y a los restantes 7 no recibieron el tratamiento. Luego de un tiempo se midió el aumento de peso diario de los terneros. Los datos obtenidos son los siguientes:

	Aumento de peso [kg/día]
Terneros tratados	0,7 - 0,7 - 0,6 - 0,7 - 0,6
Terneros no tratados	0,6 - 0,6 - 0,4 - 0,5 - 0,4 - 0,5 - 0,5

(handwritten: $\bar{x}$: 0,66; 0,5 — s: 0,055; 0,081)

a. *Identificar las unidades muestrales, las muestras y las poblaciones..*

b. *Formular y poner a prueba hipótesis para evaluar si la aplicación de cobre resulta en un aumento de la ganancia de peso de los terneros.*

c. *¿Qué tipo de error podría haber cometido? Descríbalo en términos de este problema.*

d. *Construir el intervalo de confianza correspondiente.*

e. *¿Bajo qué supuestos el válida la inferencia realizada en a. y c.?*

7.12 La aptitud de la harina de trigo para panificación depende principalmente de su contenido de un complejo proteico denominado gluten. Para evaluar la posible influencia del sistema de secado del grano sobre su contenido gluten, se seleccionaron al azar en la provincia de Buenos Aires 7 plantas de acopio que utilizan un sistema de secado prolongado a baja temperatura y 9 plantas de acopio que utilizan un sistema de secado rápido con alta temperatura y se determinó el contenido de gluten del trigo (g gluten/ 100 g harina) procesado en cada una. Los datos obtenidos son los siguientes:

	N	$\bar{x}$ [g/100g]	s^2[g/100g]^2
Baja Temp	5	25,753	1,754
Alta Temp	7	23.923	1,597

a. *Identificar las unidades muestrales, las muestras y las poblaciones*

b. *Formular hipótesis apropiadas para evaluar si el contenido esperado de gluten es afectado por el sistema de secado.*

c. *Poner a prueba la hipótesis nula con un nivel de significación α = 0,05.*

d. *Explicar que conclusión debería extraerse.*

ANÁLISIS DE LA ASOCIACIÓN ENTRE DOS VARIABLES

Hasta ahora, hemos estado tratando con muestras en las cuales se registraban o medían los valores de una variable aleatoria. Sin embargo, la mayor parte de los problemas en la ciencia y la técnica involucran más de una variable y en las muestras que se toman con el fin de analizar estadísticamente un problema o para tratar de contestar una pregunta en términos probabilísticos, se registran o miden varias variables. En esta clase sólo veremos el caso en que se registran dos variables.

Hay dos tipos básicos de problemas:

1. ambas variables son aleatorias, es decir, que en las unidades que componen las muestras que se toman aleatoriamente se miden dos variables que denotaremos X e Y – este tipo de muestras se llaman **muestras bivariadas** – y no existe ninguna relación de dependencia clara entre ambas variables aleatorias, y,

2. una de las variables (Y), llamada variable **respuesta** o variable **dependiente**, es una variable aleatoria claramente dependiente de la otra (X) a la que se llama variable **predictora** o **independiente**, que asume valores fijos dictados por el ingeniero o el experimentador.

Para analizar el primer tipo de problema, utilizaremos dos técnicas estadísticas denominadas análisis de **correlación** y análisis de **regresión**; para el segundo utilizaremos el análisis de **regresión**. En un caso como éste en el que sólo tratamos con dos variables, la regresión se dice que es **simple** y dado que sólo utilizaremos funciones lineales elementales para describir el tipo de relación entre X e Y, la técnica que utilizaremos será la del análisis de **regresión lineal simple**.

El concepto de covariancia

Consideremos el siguiente ejemplo de una muestra bivariada, donde X es el contenido de un micronutriente en el suelo (en ppm) e Y es contenido de un macronutriente (en ppm), para un grupo de muestras de suelo:

X	0.01	0.18	0.23	0.29	0.47	0.59	0.88	0.99	1.06	1.38
Y	55.2	59.9	77.3	79.0	92.1	118.3	121.5	129.4	152.7	144.6

Los datos de una muestra bivariada pueden ser gráficamente representados en un representados en un **diagrama de dispersión** como el que se muestra en la Figura 8.1. En este caso, el diagrama de dispersión mostrado permite observar que existe una asociación positiva entre las dos variables (*cuando aumenta X también aumenta Y*).

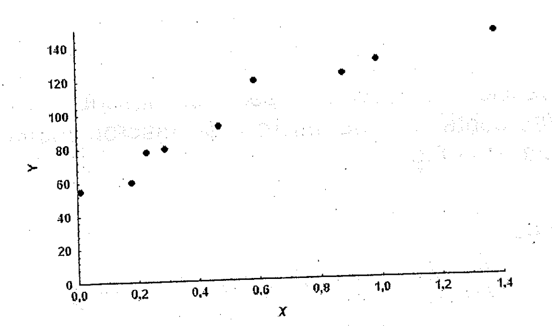


Figura 8.1. Diagrama de dispersión.

Así como existen medidas de **tendencia central** (medias, medianas, etc.) y de **dispersión** (variancia, desvío standard, coeficientes de variación, etc.) para describir la distribución de una variable aleatoria, también existen medidas que sirven para describir la asociación entre dos variables o, más específicamente, la manera en que dos variables aleatorias varían en forma conjunta. La medida principal del tipo de asociación entre dos variables aleatorias se denomina **covariancia** entre las variables *X* e *Y*, y se denota *Cov(X,Y)*. La *Cov(X,Y)* se calcula de la siguiente manera:

$$Cov(X,Y) = E[(X - \mu_X) \cdot (Y - \mu_Y)] \quad (8.1)$$

donde μ_X es la media de *X*, μ_Y es la media de *Y*, y $E(X \cdot Y)$ es la esperanza de los productos $X \cdot Y$.
Para el caso de una muestra aleatoria bivariada de tamaño *n*, la covariancia se estima como:

$$\frac{\sum (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{n-1} \quad (8.2)$$

Y así como existe el *coeficiente de variación* como medida de dispersión relativa independiente de las unidades de medición, también existe una medida relativa de la asociación estadística entre dos variables que es, también, independiente de las unidades de medición, que se denomina coeficiente de **correlación**. Para el caso de una población, el **coeficiente de correlación poblacional** entre dos variables es un parámetro que se denota con la letra ρ y que se define de la siguiente manera:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X) \cdot V(Y)}} \quad (8.3)$$

Para el caso de una muestra bivariada de tamaño *n*, estimamos el coeficiente de correlación mediante el **coeficiente de correlación muestral** que se denota mediante la letra *r* y se calcula de la siguiente manera:

$$r = \frac{\sum (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \cdot \sum (Y_i - \bar{Y})^2}} \quad (8.4)$$

La covariancia puede tomar cualquier valor en la escala de los números reales, y tanto valores positivos como negativos mientras que el coeficiente de correlación, por su naturaleza relativa, sólo puede tomar valores en el intervalo [-1,+1]. Ambas medidas, cuando son positivas, describen una asociación de tipo directo entre las variables (es decir, cuando aumenta una de ellas, la otra también tiende a aumentar) mientras que cuando son negativas, describen una asociación de tipo inverso entre las variables (es decir, cuando aumenta una de ellas, la otra tiende a disminuir).

Ejemplos

1. Los coeficientes descriptos permiten describir la asociación positiva entre los contenidos del micro y del macronutriente del suelo que se visualiza en la Figura 8.1.

   Entonces:

$$\frac{\sum (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{n-1} = \frac{135.30}{9} = 15.03 \text{ y}$$

$$r = \frac{\sum (X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \cdot \sum (Y_i - \overline{Y})^2}}$$

$$= \frac{135.30}{141.479}$$

$$= 0.9563$$

En la Tabla 8.1 se presentan los cálculos.

Tabla 8.1

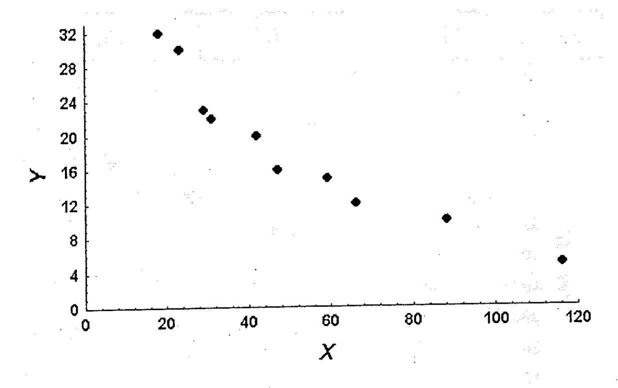
1	0.01	55.2	-0.598	-47.8	0.3576	2284.84	28.584
Obs.	X	Y	$(X_i - \overline{X})$	$(Y_i - \overline{Y})$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})^2$	$(X_i - \overline{X}) \cdot (Y_i - \overline{Y})$
2	0.18	59.9	-0.428	-43.1	0.1832	1857.61	18.447
3	0.23	77.3	-0.378	-25.7	0.1429	660.49	9.715
4	0.29	79	-0.318	-24	0.1011	576	7.632
5	0.47	92.1	-0.138	-10.9	0.0190	118.81	1.504
6	0.59	118.3	-0.018	+15.3	0.0003	234.09	-0.275
7	0.88	121.5	+0.272	+18.5	0.0740	342.25	5.032
8	0.99	129.4	+0.382	+26.4	0.1459	696.96	10.085
9	1.06	152.7	+0.452	+49.7	0.2043	2470.09	22.464
10	1.38	144.6	+0.772	+41.6	0.5960	1730.56	32.115
Suma	6.08	1030	0	0	1.8244	10971.7	**135.303**

2. Sea la siguiente muestra bivariada donde X es el número de horas de frío recibidas por un grupo de semillas e Y es el número de días transcurridos desde siembra a germinación:

X	18	23	29	31	42	47	59	66	88	116
Y	32	30	23	22	20	16	15	12	10	5

El gráfico de dispersión correspondiente se presenta en la Figura 8.2.

Figura 8.2. Diagrama de dispersión.



En este caso, puede visualizarse que la asociación entre las dos variables es negativa. Esto concuerda con los valores negativos de la covarianza y del coeficiente de correlación estimados:

$$\frac{\sum (X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{n-1} = \frac{-2257.5}{9} = -250.833 \text{ y}$$

$$r = \frac{\sum (X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \cdot \sum (Y_i - \overline{Y})^2}} = \frac{-2257.5}{2408.39} = -0.9373$$

Regresión lineal simple

Podemos distinguir dos tipos principales de relación entre variables:

1) relación **funcional** y,
2) relación **estadística**.

La primera puede ser expresada por una fórmula o modelo matemático. Es el caso de la relación entre el costo de un traslado de mercadería (*Y*) y la distancia a recorrer (*X*), cuando el costo fijo por el traslado es de $30 y se suman $5 por cada km de recorrido. En este caso el costo total del traslado se puede calcular exactamente mediante la siguiente función:

$$Y = 30 + 5 \cdot X$$

Se trata de una función que representa a una línea recta, donde la ordenada al origen es 30 (precio que nos cobra el flete sólo por haber sido contratado y llegar al lugar de partida, aunque luego decidamos no realizar el transporte) y la pendiente es 5 (incremento del costo por cada km de aumento del recorrido). Si se desea calcular el costo de un traslado a 6 km, basta con reemplazar en la función, la variable X por el valor 6 y realizar la cuenta, para enterarnos que deberemos pagar $60. En la figura 8.3 se observa que todos los puntos que satisfacen la relación se encuentran sobre la misma línea recta y que a cada valor de ***X*** le corresponde un único valor de ***Y***.

A diferencia de la relación funcional, la relación estadística no es una relación perfecta. En general, las observaciones no caen directamente sobre una línea recta. Por ejemplo, si se estudia el tiempo hasta floración de una especie, en función de la temperatura, se puede obtener una muestra de datos como la siguiente, que representa la suma de temperaturas (X) por encima de un umbral y los días hasta floración (Y) para la especie en cuestión

X	18	23	31	31	42	59	19	78	47	59	66	23	36	66	60
Y	32	30	25	22	20	12	26	7	16	15	12	26	18	9	9

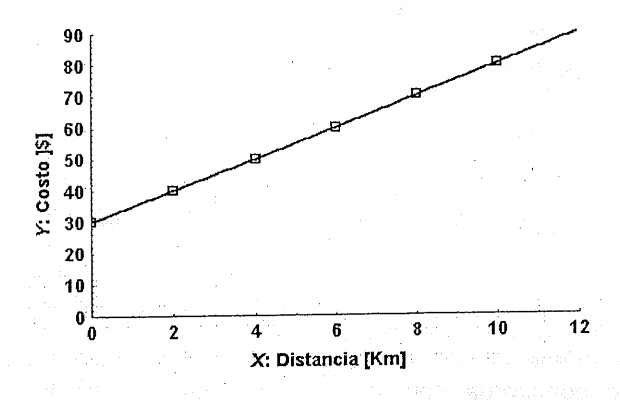
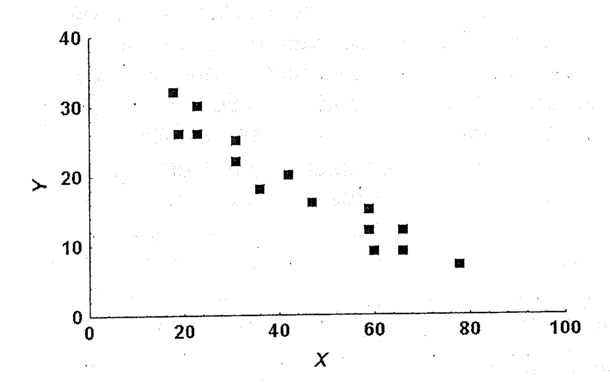


Figura 8.3.

El diagrama de dispersión correspondiente a esta muestra, que se presenta en la Figura 8.4, sugiere que hay claramente una relación lineal entre la

suma de temperaturas y el *tiempo hasta floración*, en el sentido de que a mayor temperatura, la floración ocurre más temprano.

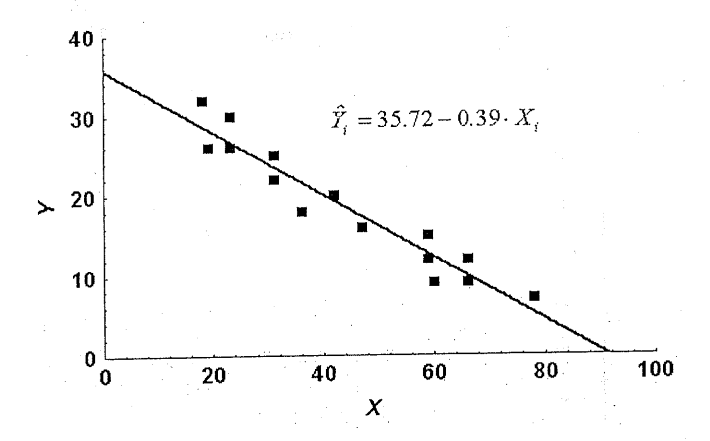
Figura 8.4. Diagrama de dispersión.



Sin embargo, puede verse que la relación no es perfecta: para cada valor de suma de temperaturas no existe un único tiempo hasta floración, sino que hay una dispersión de puntos sugiriendo que parte de la variación en el tiempo hasta floración no se explica por la suma de temperaturas. En este caso el tiempo hasta floración es la **variable dependiente** o **variable respuesta** (*Y*) y la suma de temperaturas, la **variable independiente** o **variable predictora** (*X*).

La Figura 8.5 muestra la recta que describe la relación estadística entre las variables estudiadas (luego explicaremos como obtenerla). La dispersión de puntos alrededor de la línea representa la variación en tiempo a floración que no está asociada linealmente a la suma de temperaturas.

Figura 8.5. Diagrama de dispersión con recta de dispersión.



La técnica de análisis de regresión lineal simple se utiliza para analizar la *relación estadística* entre dos variables. Debe quedar claro desde ahora que la relación entre las dos variables que se pretende determinar es de naturaliza *estadística* y no solamente matemática, siempre habrá un grado de incertidumbre en cuanto a las relaciones que se establezcan y en cuanto a las estimaciones y pruebas de hipótesis que se hagan.

Emplearemos la relación funcional más simple: la línea recta que queda completamente definida una vez conocidos su *ordenada al origen* y su *pendiente*. El objetivo de la técnica consiste en encontrar la línea recta que mejor describa la relación entre las variables predictora (*X*) y respuesta (*Y*).

Ejemplo:
El ejemplo consiste en 10 lotes de *Picea* en un gran vivero de Bariloche. En dicha muestra se midieron dos variables: el *tamaño del lote* de producción y el *número de Horas-Hombre* insumidas para producir arbolitos en dicho lote.

El número de Horas-Hombre es la **variable dependiente** o **variable respuesta** (*Y*) y el tamaño del lote, la **variable independiente** o **variable predictora** (*X*). En la Tabla 8.2 se muestran los datos. La Figura

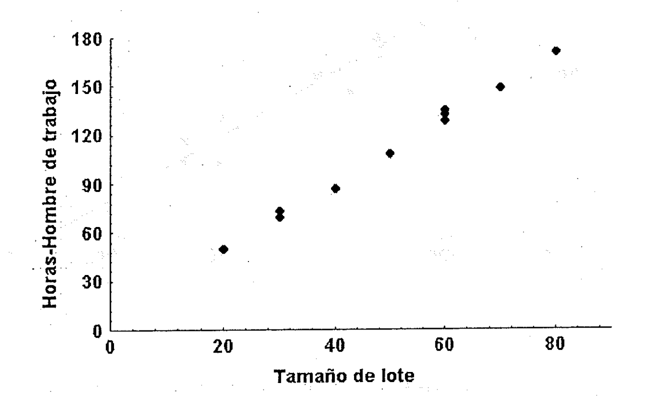
8.6 muestra la dispersión de los datos. Se nota claramente en el gráfico que la relación entre las dos variables es directa: a mayor tamaño de lote, mayor cantidad de Horas-hombre necesarias para producirlo.

De conocer todos los valores posibles de ambas variables (estaríamos tratando con una población), entonces se podría buscar una recta que describa ajustadamente la relación entre las dos variables, es decir que, si se hallara dicha recta, se conocerían sus parámetros: la ordenada al origen (β_0) y la pendiente (β_1). Pero si eso no es posible, solo se podrá disponer de los datos de una muestra. La cuestión ahora es encontrar la recta que mejor "ajuste" los puntos del diagrama de dispersión, es decir que, a partir de los datos de la muestra se deberán encontrar estimadores de los parámetros β_0 y β_1 de la recta verdadera (en la población) a los que denotaremos como b_0 y b_1.

Tabla 8.2.

Salida de producción (i)	Tamaño de lote (X_i)	Horas -Hombre (Y_i)
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132

Figura 8.6.



No esperaremos que todos los puntos muestrales caigan exactamente sobre ella sino que habrá una diferencia debida al *error* de la muestra. Para expresar la relación estadística entre las dos variables tendremos que escribir el **modelo de regresión**:

Modelos de regresión

Un **modelo de regresión** es una manera formal de expresar los dos ingredientes esenciales de una relación estadística:

- una tendencia de la variable dependiente Y a variar conjuntamente con la variación de la (o las) variable(s) independiente(s) de una manera sistemática y,
- una dispersión de las observaciones alrededor de la curva de la relación estadística.

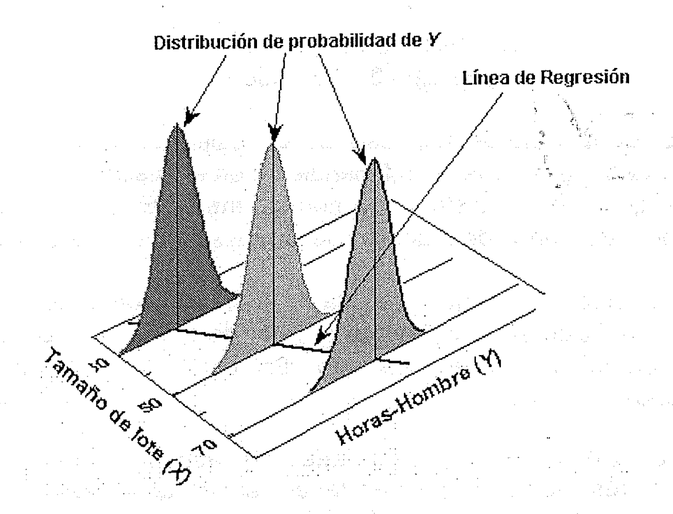
Estas dos características están implícitas en un modelo de regresión postulando que:

- en la población de observaciones asociadas con el proceso que fue muestreado, hay una distribución de probabilidades de *Y* para cada nivel de *X*.

- las medias de estas distribuciones de probabilidades varían de una manera sistemática al variar *X*.

Siguiendo con el ejemplo, para cada tamaño de lote, se asume que hay una distribución de probabilidades de *Y*. La Figura 7 muestra esa distribución para *X* = 30 que es el tamaño de lote para la primera salida de producción. Entonces, la cantidad real de Horas-Hombre (73) es vista como una selección aleatoria a partir de esta distribución de probabilidades.

Figura 8.7.



La Figura 8.7 también muestra las distribuciones de probabilidades de *Y* para los tamaños de lote 50 y 70 (*X* = 50 y *X* = 70). Nótese que las medias de las distribuciones de probabilidades guardan una relación exacta con el nivel de *X*. Esta relación exacta se denomina **función de regresión de *Y* sobre *X***. El gráfico de la función de regresión se denomina **curva de regresión**. En la figura la función de regresión es lineal. Para nuestro ejemplo, esto implicaría que el número esperado (es decir, la media) de Horas-Hombre varía de manera lineal con la variación en el tamaño del lote. El número de Horas-Hombre podría estar relacionado de otra manera con el tamaño del lote - no necesariamente deberá ser una línea recta – pero en este curso sólo estudiaremos relaciones lineales.

Objetivos del análisis de regresión

El análisis de regresión persigue tres grandes objetivos: (1) *descripción*, (2) *control* y, (3) *predicción*.
En los estudios **observacionales**, es decir, cuando se observa un proceso sin incidir sobre el mismo (o tratando de no hacerlo) el propósito es claramente descriptivo. Por ejemplo, en el estudio de la influencia de la cantidad de dióxido de azufre en el aire (*X*) sobre el porcentaje de plantas atacadas por un insecto en un bosque (*Y*), se tomarán muestras bivariadas y se registrarán los valores de ambas variables con el fin de describir ese proceso de contaminación-infestación.
En los estudios **técnicos** donde el ingeniero manipula una variable (*X*) y observa cómo cambia otra (*Y*), el propósito es controlar el proceso con fines técnicos o económicos. Por ejemplo, la manipulación de dosis de fertilizantes sobre el rendimiento de un cultivo: un ensayo permitiría hallar una relación estadística entre rendimientos y dosis de fertilizante en el cultivo para fijar los gastos en ese rubro. Finalmente, conocer la relación estadística funcional entre dos variables permite predecir el comportamiento futuro de una de ellas dado que se conoce el

valor de la otra. Por ejemplo, si se conoce la relación funcional que liga el porcentaje de humedad en el ambiente con el porcentaje de plantas infectadas por un hongo patógeno en un cultivo, se puede lanzar una alarma cuando el nivel de humedad llegue a un nivel crítico. Pero los distintos objetivos suelen superponerse. El ejemplo de los tamaños de lotes y las Horas-Hombre permite la predicción del requerimiento en Horas-Hombre para la próxima salida de producción dado un tamaño de lote, a los fines de la estimación de los costos y la programación de la producción. Después que la producción se completó, los ingenieros pueden comparar las Horas-Hombre reales con las horas predichas por el modelo a los fines del control administrativo.

Modelo de regresión lineal

El modelo básico del que hemos venido hablando puede formalizarse de la siguiente manera:

$$\begin{aligned} &Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i \\ &\varepsilon_i \sim Normal\,(0, \sigma^2) \\ &Cov(\varepsilon_i; \varepsilon_j) = 0 \quad \text{cuando } i \neq j \end{aligned} \tag{8.5}$$

donde Y_i es el valor de la variable respuesta en el i-ésimo ensayo, β_0 y β_1 son parámetros, X_i es el valor de la variable independiente en el i-ésimo ensayo y ε_i es un término de error aleatorio con distribución normal, media $E(\varepsilon_i) = 0$ y variancia σ^2; ε_i y ε_j no están correlacionados de manera que $Cov(\varepsilon_i; \varepsilon_j) = 0$ para todas las i y j, con $i \neq j$ e $i = 1,2,\ldots,n$.

Como puede verse el valor de Y_i resulta de sumar un componente exacto determinado por los coeficientes β_0 y β_1 y por el valor de X_i y un componente no exacto o aleatorio determinado por el valor de ε_i. Por este motivo, el valor de Y_i también será aleatorio y, como tal:

1. tendrá una distribución de probabilidades y, puesto que hemos asumido que los errores aleatorios pueden tener valores tanto positivos como negativos con media total igual a 0, dicha distribución de probabilidades tendrá media igual a: $E(Y_i) = E(\beta_0 + \beta_1 \cdot X_i + \varepsilon_i) = \beta_0 + \beta_1 \cdot X_i + E(\varepsilon_i) = \beta_0 + \beta_1 \cdot X_i$, es decir, el valor de la función de regresión lineal y la diferencia entre esa media y el valor observado (Y_i) y $\beta_0 + \beta_1 \cdot X_i$ es, justamente, el valor del error correspondiente a esa unidad (ε_i);

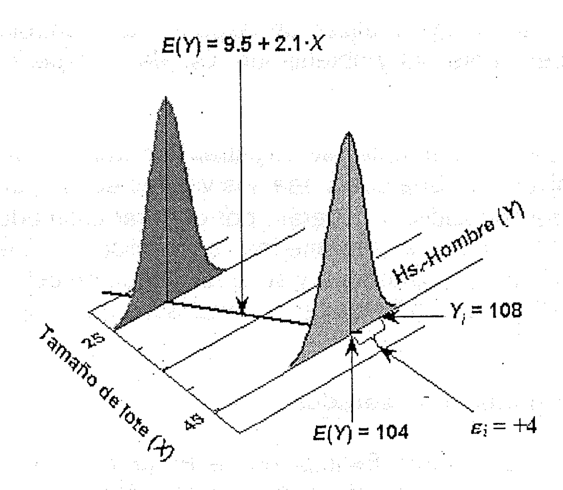
2. puesto que la variancia de los ε_i es igual a σ^2, $V(Y_i) = V(\beta_0 + \beta_1 \cdot X_i + \varepsilon_i) = 0 + V(\varepsilon_i) = \sigma^2$ para cualquier nivel de X, y

3. puesto se supone que los ε_i son independientes, también se supone que los diferentes resultados obtenidos, Y_i, son completamente independientes (es decir que el valor de uno de ellos no tiene ninguna influencia sobre el valor de otro de ellos).

Siguiendo con el ejemplo, supongamos que un modelo de regresión lineal se puede aplicar al ejemplo de los tamaños de lote y que dicho modelo es: $Y_i = 9.5 + 2.1 \cdot X_i + \varepsilon_i$. La siguiente figura contiene una representación de la función de regresión $E(Y) = 9.5 + 2.1 \cdot X$. Supongamos que en la i-ésima unidad se produce un lote de $X_i = 45$ unidades y que el número observado de Horas-Hombre es $Y_i = 108$. En este caso, el término del error es $\varepsilon_i = +4$ porque $E(Y_i) = 9.5 + 2.1 \cdot (45) = 104$ e $Y_i = 108 = 104 + 4$.

La Figura 8.8 muestra la distribución de probabilidad de Y cuando $X = 45$ e indica dónde está la observación $Y_i = 108$ en esta distribución. Nótese otra vez que el término del error ε_i es, simplemente, la desviación de la observación con respecto a su valor promedio $E(Y_i)$. La figura también muestra la distribución de probabilidad de Y cuando $X = 25$. Nótese que esta distribución muestra la misma variabilidad que la distribución de probabilidad correspondiente a $X = 45$, de conformidad con los requerimientos del modelo lineal simple.

Figura 8.8.

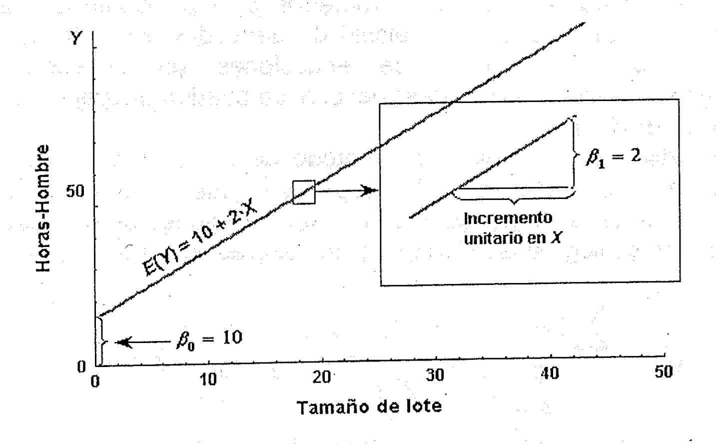


Parámetros de la regresión

Los parámetros β_0 y β_1 se denominan **coeficientes de regresión**. β_1 es la **pendiente** de la **línea de regresión** e indica el cambio en la media de la distribución de probabilidad de Y por cada unidad de incremento en X. El parámetro β_0 es la ordenada al origen (intercepción) de la línea de regresión. Si el rango de valores del modelo llega hasta $X = 0$, β_0 da la media de la distribución de probabilidad de Y en $X = 0$. Cuando el rango del modelo no llega hasta $X = 0$, β_0 no tiene ningún significado particular como término en el modelo de regresión.

Ejemplo.
La Figura 8.9 muestra la función de regresión $E(Y) = 10 + 2 \cdot X$ para el ejemplo anterior de los tamaños de lotes. La pendiente $\beta_1 = 2$ indica que un incremento de una unidad en el tamaño del lote lleva a un incremento en la media de la distribución de probabilidad de Y de 2 Horas-Hombre. La ordenada al origen $\beta_0 = 10$ indica el valor de la función de regresión en $X = 0$, pero como el modelo de regresión lineal fue formulado para que se aplique a tamaños de lote que iban desde 20 hasta 80 unidades, β_0 no tiene ningún significado por sí mismo y, en particular, no indica necesariamente el tiempo promedio al comienzo del proceso, es decir el número promedio de Horas-Hombre antes de que comience la producción.

Figura 8.9.



Estimación del Modelo de regresión

Como hemos dicho antes, se puede realizar un experimento controlando los valores de la variable independiente (X_i) y obteniendo, así, datos *experimentales*, o un

estudio *observacional* donde, simplemente, se registran los valores de ambas variables en una muestra bivariada. Sea como sea, los valores de los parámetros β_0 y β_1 serán, en general, desconocidos y deberán, por ello, ser estimados. En la clase correspondiente a Estimación de Parámetros, se explicó un método de estimación (el método de máxima verosimilitud) y se anunció que en esta clase de Regresión Lineal se explicaría el otro (el método de cuadrados mínimos). Aquí lo haremos.

Método de estimación por mínimos cuadrados

Tal como se indicara en la clase sobre Estimación de Parámetros, otro de los métodos de estimación que vemos en este curso es el método de **mínimos cuadrados**. Se supone que las observaciones de la muestra tienen la forma (para el caso de un parámetro único, θ):

$$Y_i = f_i(\theta) + \varepsilon_i, \qquad i = 1,2,\ldots,n \qquad (8.6)$$

donde $f_i(\theta)$ es una función conocida del parámetro θ y las ε_i son variables aleatorias de las cuales se asume, comúnmente, que tiene esperanza igual a 0, es decir, $E(\varepsilon_i) = 0$. Con el método de mínimos cuadrados, para un conjunto de observaciones mueestrales dado, la suma de cuadrados:

$$Q = \sum_{i=1}^{n} \left[Y_i - f_i(\theta)\right]^2 \qquad (8.7)$$

es considerada como una función de θ. El estimador de mínimos cuadrados de θ se obtiene minimizando Q con respecto a θ, es decir, derivando Q con respecto a θ e igualando a 0. En muchas instancias, los estimadores de mínimos cuadrados son insesgados y consistentes.Este método utiliza los cuadrados de las diferencias entre las observaciones Y_i y sus valores esperados:

$$Q = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 \cdot X_i\right)^2 \qquad (8.8)$$

Y buscará los valores b_0 y b_1 que hagan que Q tenga su valor mínimo: ésos serán los estimadores de los parámetros β_0 y β_1. Como es sabido, para hallar mínimos se debe recurrir al cálculo de derivadas. En este caso que nos ocupa, tendremos un sistema de ecuaciones en derivadas parciales (denominadas **ecuaciones normales**) del cual se pueden despejar los valores de b_0 y b_1: (VER ANEXO I)

Como dijimos, el objetivo del método de mínimos cuadrados es hallar estimaciones b_0 y b_1 para β_0 y β_1, respectivamente, para las cuales **Q sea mínima**. Después de las correspondientes manipulaciones algebraicas (que aquí no detallaremos), se llega a las siguientes dos expresiones para b_0 y b_1:

$$b_1 = \frac{\sum \left(X_i - \overline{X}\right)\cdot\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2} \quad \text{y} \quad b_0 = \overline{Y} - b_1 \cdot \overline{X} \qquad (8.9)$$

donde $\overline{X}$ e $\overline{Y}$ son las medias de X e Y, respectivamente.

Ejemplo.
Para ilustrar el cálculo de los estimadores de mínimos cuadrados b_0 y b_1, utilizaremos, nuevamente, el ejemplo de los tamaños de lotes en el vivero

de *Picea* en Bariloche, cuyos datos muestrales se presentaron y se graficaron en la página 89. Los cálculos se presentan en la Tabla 8.3.

Tabla 8. 3

X	Y	$(X_i-\overline{X})^2$	$(Y_i-\overline{Y})^2$	$(X_i-\overline{X})\cdot(Y_i-\overline{Y})$
30	73	400	1369	740
20	50	900	3600	1800
60	128	100	324	180
80	170	900	3600	1800
40	87	100	529	230
50	108	0	4	0
60	135	100	625	250
30	69	400	1681	820
70	148	400	1444	760
60	132	100	484	220
Suma		**3400**	**13660**	**6800**

SC_X = 3400, SC_Y = 13660 y SP_{XY} = 6800.

Luego: $$b_1=\frac{\sum(X_i-\overline{X})\cdot(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2}=\frac{SP_{XY}}{SC_X}=2.0$$

$$b_0=\overline{Y}-b_1\cdot\overline{X}=110-2\cdot(50)=10,$$

donde $\overline{X}$ = 50 e $\overline{Y}$ = 110 son las medias de X e Y, respectivamente.

SC_Y es una medida de la **variación total de la respuesta** y su utilidad se verá más adelante.

Obtenemos, b_0 = 10 y b_1 = 2. Así, estimamos que el número medio de Horas-Hombre aumenta en 2.0 horas por cada unidad de incremento en el tamaño del lote, como indica la pendiente b_1 = 2.0. La ordenada al origen b_0 = 10 indica el valor de la función de regresión en X = 0, pero como el modelo de regresión lineal fue formulado para que se aplique a tamaños de lote que iban desde 20 hasta 80 unidades, β_0 (y, por lo tanto b_0) no tiene ningún significado por sí mismo y, en particular en este ejemplo, no indica el tiempo promedio para lotes de dimensión igual a cero.

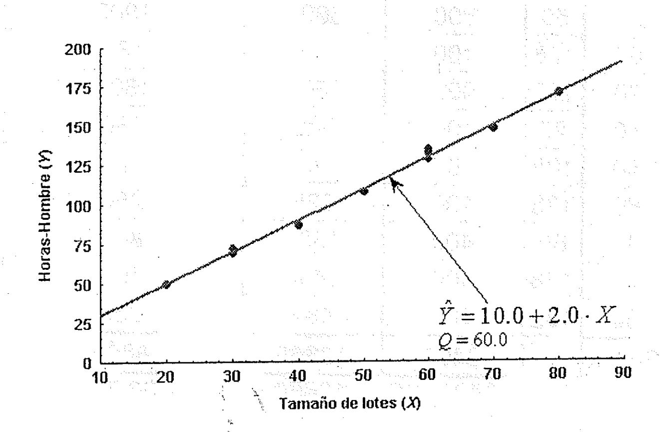
Estimación de la media de *Y* dado *X*

Los estimadores de β_0 y β_1, respectivamente b_0 y b_1, pueden ser usados para estimar los valores de la media de Y correspondientes a valores dados de la variable independiente X usando la fórmula $\hat{Y}=b_0+b_1\cdot X$, donde el signo sobre la Y se lee "estimado" o "ajustado" y es el valor de la función de regresión correspondiente a un valor de X. La diferencia entre un valor observado y el correspondiente valor ajustado por la recta de regresión se denomina **residual** de dicha observación: $e_i=Y_i-\hat{Y}_i$. En el cuadro de la izquierda presentamos los cálculos correspondientes al ejemplo que venimos utilizando.

En este caso de los tamaños de lotes, hallamos que las estimaciones de mínimos cuadrados de los coeficientes de regresión eran b_0 = 10.0 y b_1 = 2.0; por tanto, la función de regresión estimada es $\hat{Y}=10.0+2.0\cdot X$. Si estamos interesados en el número medio de Horas-Hombre cuando el tamaño de lote es, por ejemplo, X = 55, nuestra estimación puntual sería $\hat{Y}=10.0+2.0\cdot 55=120$. Así, estimaríamos que el número medio de Horas-Hombre para los lotes de tamaño X = 55 es igual a 120. Esto significa que si se producen muchas tandas

con lotes de tamaño 55 bajo las condiciones de las 10 tandas de la muestra, el tiempo de trabajo promedio para cada tanda será de alrededor de 120 horas. Desde ya que el tiempo de trabajo para un lotea de tamaño 55 cualquiera es probable que sea más alto o más bajo que la respuesta media debido a la variabilidad inherente en el sistema, tal como se representa mediante el término del error en el modelo. La Figura 8.10 contiene un gráfico de la función de regresión estimada $\hat{Y} = 10.0 + 2.0 \cdot X$, así como los datos originales.

Figura 8.10



Los valores ajustados para los datos muestrales son obtenidos sustituyendo los valores de X de la muestra en la ecuación de regresión estimada. Por ejemplo, para los datos de la muestra del ejemplo, $X_1 = 30$. Por tanto, el valor ajustado es: $\hat{Y} = 10.0 + 2.0 \cdot 30 = 70$. Esto se compara con el valor observado de Horas-Hombre, $Y = 73$. La Tabla 4 contiene los valores de la variable independiente (X_i), las respuestas (Y_i), los valores ajustados por el modelo de regresión lineal ($\hat{Y}_i$), los residuales y sus cuadrados.

Residuales

El i-ésimo **residual** es la diferencia entre el valor observado Y_i y el correspondiente valor ajustado $\hat{Y}_i$:

$$e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 \cdot X_i .$$

La Figura 8.11 muestra los 10 residuales del ejemplo. Las magnitudes de los residuales se muestran mediante líneas verticales entre cada observación y el valor ajustado sobre la línea de regresión estimada. Debemos distinguir entre el valor del término del **error** del modelo, $\varepsilon_i = Y_i - E(Y_i)$, y el **residual**, $e_i = Y_i - \hat{Y}_i$. El primero se refiere a la desviación vertical de Y_i con respecto a la línea de regresión poblacional desconocida y, por tanto, es desconocido. Por otra parte, el residual es la desviación vertical **observada** de Y_i con respecto a la línea de regresión ajustada.

Los residuales son muy útiles para estudiar si un modelo de regresión es apropiado para los datos con los cuales se está trabajando.

Figura 8.11

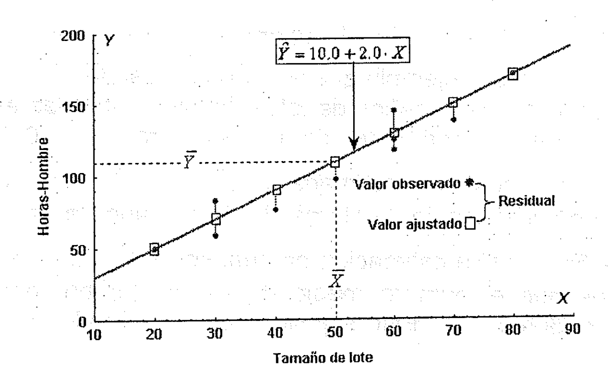


Tabla 8.4.

Observ.	Tamaño de lote (X_i)	Horas-Hombre (Y_i)	Respuesta media estimada $(\hat{Y}_i)$	Residual $e_i = Y_i - \hat{Y}_i$	Residual al cuadrado $e_i^2 = (Y_i - \hat{Y}_i)^2$
1	30	73	70	+3	9
2	20	50	50	0	0
3	60	128	130	-2	4
4	80	170	170	0	0
5	40	87	90	-3	9
6	50	108	110	-2	4
7	60	135	130	+5	25
8	30	69	70	-1	1
9	70	148	150	-2	4
10	60	132	130	+2	4
Total	500	1100	1100	0	Q = 60

Propiedades de la línea de regresión ajustada

La línea de regresión ajustada por el método de mínimos cuadrados tiene ciertas propiedades que vale la pena mencionar.

- *La suma de los residuales es igual a 0:* $\sum_{i=1}^{n} e_i = 0$ y, como consecuencia de esta propiedad, tenemos la propiedad de que *la suma de los valores observados Y_i es igual a la suma de los valores ajustados,* $\hat{Y}_i$: $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$.
- *La suma de los residuales elevados al cuadrado,* $\sum e_i^2$ *, es un mínimo.*
- *La línea de regresión siempre pasa por el punto* $(\bar{x}; \bar{y})$.

Estimación de la variancia del error (σ^2)

La variancia del error, σ^2, es también una medida de qué tan bueno es el ajuste realizado por la función de regresión. Es necesario tener una estimación de la variancia del error a partir de los datos de la muestra.

Para poder obtener una estimación de σ^2, es necesario conocer los valores de los residuales del análisis de regresión, $Y_i - \hat{Y}_i = e_i$, y obtener la suma de sus cuadrados, que denotaremos SC_E:

$$SC_E = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 \cdot X_i)^2 = \sum_{i=1}^{n} e_i^2 \qquad (8.10)$$

Finalmente, calcularemos la variancia correspondiente a dicha suma de cuadrados – que se denomina **cuadrado medio del error** y que denotaremos CM_E - dividiéndola por sus grados de libertad:

$$
\begin{aligned}
CM_E &= \frac{SC_E}{n-2} \\
&= \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2} \\
&= \frac{\sum (Y_i - b_0 - b_1 \cdot X_i)^2}{n-2} \\
&= \frac{\sum e_i^2}{n-2}
\end{aligned}
\qquad (8.11)
$$

y éste es el estimador de la variancia del error que estamos buscando: $E(CM_E) = \sigma^2$.

Para realizar las inferencias necesarias para tomar decisiones, debemos suponer una distribución para los términos del error. Para el modelo que estamos utilizando supondremos que los errores tienen distribución normal con media igual a 0 y variancia igual a σ^2, es decir que el modelo de regresión completo es el siguiente: $Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$, donde Y_i es el valor de la variable respuesta correspondiente a la i-ésima unidad, X_i es el valor de la variable independiente en esa misma unidad, β_0 y β_1 son los parámetros de la regresión y los ε_i son los errores independientes que tienen distribución normal con media 0 y variancia σ^2.

Coeficiente de determinación

El coeficiente de determinación, R^2, es una medida descriptiva del grado de asociación lineal entre las dos variables. Está compuesto por la Suma de Cuadrados Total (SC_{TOT}), que mide la variación total en las observaciones Y_i, y la Suma de Cuadrados de Error (SC_E) que mide la variación residual en las Y_i cuando se emplea el modelo de regresión. Una medida natural de la magnitud del efecto de X de reducir la variación en Y es:

$$R^2 = \frac{SC_{TOT} - SC_E}{SC_{TOT}} = \frac{SC_R}{SC_{TOT}} \qquad (8.12)$$

donde:

$$SC_{TOT} = SC_Y = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (8.13)$$

$$\text{y } SC_E = \sum_{i=1}^{n} (e_i)^2 \qquad (8.14)$$

que es lo mismo que:

$$SC_E = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 = \sum_{i=1}^{n} e_i^2 \qquad (8.15)$$

El coeficiente R^2 indica la proporción de la variación total de Y puede ser explicada por la dependencia lineal de X. Entonces, $0 \leq R^2 \leq 1$. Si todas las observaciones caen en la recta ajustada y ésta no es horizontal, entonces la $SC_E = 0$ y $R^2 = 1$. La variable X explica toda la variación en las observaciones Y_i. La variación en Y está completamente ligada a X, por lo tanto, al cambiar X, cambia también Y, de tal forma que todos los puntos $(x;y)$ se ubican sobre una recta.

Si no existe regresión lineal, $R^2 = 0$, $SC_E = SC_{TOT}$, lo que indica que no hay asociación lineal entre X e Y y que la variación en X no es de ninguna ayuda para explicar la variación de las observaciones Y_i. Es decir que los valores de Y cambian en forma totalmente aleatoria con respecto a X o forman otro tipo de

asociación que no es lineal simple. En la práctica no es probable que R^2 sea exactamente igual a 0 o a 1: lo más común es que se encuentre entre ambos valores. Cuanto más cerca de 1 esté el valor, más grande será el grado de asociación lineal entre X e Y. Así, un valor de $R^2 = 0.80$ está indicando que el 80 % de la variabilidad en Y es explicada por la dependencia lineal de Y con respecto a X. Para el ejemplo de los lotes de *Picea*:

$$R^2 = \frac{SC_{TOY} - SC_E}{SC_{TOT}} = \frac{13660 - 60}{13660} = 0.995$$

es decir que el 99.5% de la variabilidad en el número de Horas-Hombre de trabajo, es explicada por el tamaño del lote.

Inferencias en el análisis de regresión

Inferencias para β_1

Como se dijo antes, β_1 es la pendiente de la línea de regresión y obtener una estimación de este parámetro nos permite tener una idea del cambio esperado en la variable respuesta ante un cambio determinado en la variable predictora. La prueba de hipótesis más común acerca de β_1 es la siguiente: H_0: $\beta_1 = 0$ vs. H_1: $\beta_1 \neq 0$. Si H_0 es cierta, entonces se estima que no existe asociación alguna entre X e Y. En la Figura 12 se muestra un caso en que $\beta_1 = 0$, es decir que $E(Y) = \beta_0 + 0 \cdot X = \beta_0$.

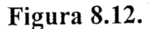
Distribución por muestreo de b_1

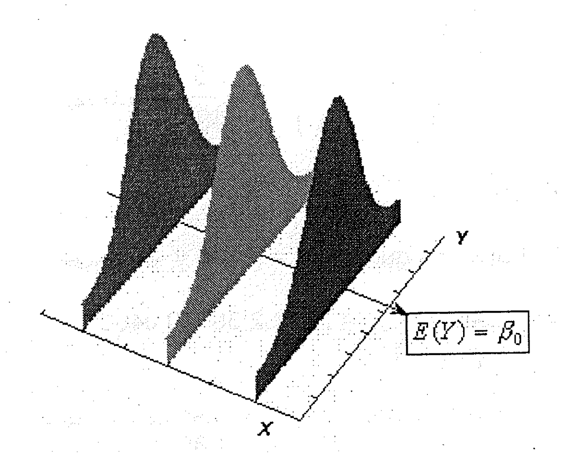
Como se adelantó al principio de la clase, el estimador puntual de β_1 es b_1:

$$b_1 = \frac{\sum (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \quad (8.16)$$

y su distribución por muestreo es normal, con las siguientes media y variancia:

$$E(b_1) = \beta_1 \text{ y } \sigma^2(b_1) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2} . \quad (8.17)$$


Figura 8.12.



Distribución por muestreo de $\dfrac{b_1-\beta_1}{s(b_1)}$

El estadístico que utilizaremos para las pruebas de hipótesis acerca de β_1 es el estadístico estandarizado $\dfrac{b_1-\beta_1}{\sigma(b_1)}$ que tiene distribución normal standard y estimaremos $\sigma(b_1)$ mediante $s(b_1)$. Finalmente, bajo H_0, $\dfrac{b_1-\beta_1}{s(b_1)}$ tiene distribución t_{n-2} para el modelo que estamos utilizando, siendo

$$s^2(b_1)=\frac{CM_E}{\sum(x_i-\bar{x})^2} \tag{8.18}$$

Intervalos de confianza para β_1

Sabido que $\dfrac{b_1-\beta_1}{s(b_1)}$ tiene distribución t, el $IC_{1-\alpha}$

Resulta: $$P\left\{t_{\alpha/2;n-2}\le\frac{b_1-\beta_1}{s(b_1)}\le t_{1-\alpha/2;n-2}\right\}=1-\alpha\text{ , o} \tag{8.19}$$

Sea: $$P\left\{b_1-t_{1-\alpha/2;n-2}\cdot s(b_1)\le\beta_1\le b_1+t_{1-\alpha/2;n-2}\cdot s(b_1)\right\}=1-\alpha \tag{8.20}$$

Ejemplo. Siguiendo con el ejemplo de los tamaños de lote, supongamos que se desea obtener un IC_{95} para β_1. Los cálculos necesarios son los siguientes:

$n = 100$; $\bar{X}=50$; $b_0 = 10.0$; $b_1 = 2.0$;
$\hat{Y}=10.0+2.0\cdot X$; $SC_E = 60$;
$CM_E = 7.5$;
$\sum(X_i-\bar{X})^2=3400$;
$\sum(X_i-\bar{X})\cdot(Y_i-\bar{Y})=6800$;
$\sum(Y_i-\bar{Y})^2=13660$;
$$s^2(b_1)=\frac{CM_E}{\sum(X_i-\bar{X})^2}=\frac{7.5}{3400}=0.002206$$

y $s(b_1) = 0.04697$.

Para el IC_{95} hallamos que $t_{8;0.975} = 2.306$ y, entonces:

$2.0 - 2.306\cdot(0.04697) \le \beta_1 \le 2.0 + 2.306\cdot(0.04697)$,
es decir, $1.89 \le \beta_1 \le 2.11$.

Así que, con una confianza del 95%, estimamos que el número medio de Horas-Hombre se incrementa entre 1.89 y 2.11 por cada incremento de una unidad en el tamaño del lote.

Pruebas de hipótesis para β_1

Prueba bilateral.
Supongamos que se desea probar si existe alguna asociación lineal entre los tamaños de los lotes y el número de Horas-Hombre, es decir: H_0: $\beta_1 = 0$; H_1: $\beta_1 \neq 0$.

La estadística de prueba es: $t^* = \dfrac{b_1}{s(b_1)}$ y la regla de decisión con un nivel de significación α es:

si $|t^*| \leq t_{1-\alpha/2;n-2}$, no se rechaza H_0; si $|t^*| > t_{1-\alpha/2;n-2}$, se rechaza H_0.

Para el ejemplo de los tamaños de lote, con $\alpha = 0.05$, $b_1 = 2.0$, $s(b_1) = 0.04697$ y $t_{8;0.975} = 2.306$ la regla de decisión es aceptar H_0 si $|t^*| \leq 2.306$ y rechazar H_0 si $|t^*| > 2.306$. Dado que:

$$|t^*| = \left|\frac{2.0}{0.04697}\right| = 42.58 > 2.306$$

se decide rechazar H_0 y concluir en que $\beta_1 \neq 0$, o sea que existe una asociación lineal entre los tamaños de los lotes y el número de Horas-Hombre. Mediante el menú ***Estadísticas – Probabilidades y cuantiles*** de ***Infostat*** podemos ver que el valor *p* para el resultado de la muestra es casi 0. Y, por tanto, el valor de *p* bilateral también es casi 0.

Prueba unilateral.

En este caso las hipótesis son: H_0: $\beta_1 \leq 0$; H_1: $\beta_1 > 0$ y la regla de decisión basada en la prueba *t*: si $|t^*| \leq t_{1-\alpha;n-2}$ se acepta H_0; si $|t^*| > t_{1-\alpha;n-2}$ se rechaza H_0. Con $\alpha = 0.05$, $t_{8;0.95} = 1.860$ y $t^* = 42.58$, decidimos rechazar H_0, o sea que concluimos en que β_1 es positivo.

Inferencias para β_0

Distribución por muestreo de b_0

Como se indicó antes, el estimador puntual b_0 es $b_0 = \overline{Y} - b_1 \cdot \overline{X}$
y la distribución por muestreo de b_0 es normal con media y variancia $E(b_0) = \beta_0$

$$\text{y} \qquad \sigma^2(b_0) = \sigma^2 \cdot \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right], \tag{8.21}$$

respectivamente.

Un estimador de $\sigma^2(b_0)$ se obtiene reemplazando σ^2 por su estimador puntual CM_E:

$$s^2(b_0) = CM_E \cdot \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right] \tag{8.22}$$

Distribución por muestreo de $\frac{b_0 - \beta_0}{s(b_0)}$.

$\frac{b_0 - \beta_0}{s(b_0)}$ tiene distribución t_{n-2}. Por tanto, se pueden establecer intervalos de confianza y pruebas de hipótesis usando la distribución *t*.

Intervalo de confianza para β_0

Límites de confianza con 1 - α para β_0: $b_0 \pm t_{n-2;1-\alpha/2} \cdot s(b_0)$.

Ejemplo. Si se desea construir un IC_{90}, hallaríamos primero $t_{8;0.95}$ y $s(b_0)$. $T_{8;0.95}$ = 1.860 y, por los resultados previos, sabemos que:

$$s^2(b_0) = CM_E \cdot \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2}\right] = 7.5 \cdot \left[\frac{1}{10} + \frac{50^2}{3400}\right] = 6.26471 \text{ y}$$

$s(b_0)$ = 2.50294.
Y el IC_{90} para β_0 es: 10.0 – 1.860·(2.50294) ≤ β_0 ≤ 10.0 + 1.860·(2.50294), es decir, 5.34 ≤ β_0 ≤ 14.66.

Inferencias para la media de *Y* dado *X*

Otro aspecto fundamental del análisis de regresión es que, conociendo la función de regresión que ajusta los datos, también se puede conocer el valor esperado de la variable respuesta, $E(Y_k)$, correspondiente a un valor determinado de la variable predictora, X_k, Por tanto, también se pueden construir intervalos de confianza con respecto a Y_k. El estimador puntual de $E(Y_k)$ es $\hat{Y}_k : \hat{Y}_k = b_0 + b_1 \cdot X_k$.

Distribución por muestreo de $\hat{Y}_k$

La distribución por muestreo de $\hat{Y}_k$ es normal con las siguientes media y variancia:

$$E(\hat{Y}_k) = E(Y_k) \text{ y } \sigma^2(\hat{Y}_k) = \sigma^2 \cdot \left[\frac{1}{n} + \frac{(X_k - \overline{X})^2}{\sum (X_i - \overline{X})^2}\right] \quad (8.23)$$

Cuando CM_E es sustituido por σ^2 se obtiene $s^2(\hat{Y}_k)$, la variancia estimada de $\hat{Y}_k$:

$$s^2(\hat{Y}_k) = CM_E \cdot \left[\frac{1}{n} + \frac{(X_k - \overline{X})^2}{\sum (X_i - \overline{X})^2}\right] \quad (8.24)$$

Distribución por muestreo de $\frac{\hat{Y}_k - E(Y_k)}{s(\hat{Y}_k)}$

$\frac{\hat{Y}_k - E(Y_k)}{s(\hat{Y}_k)}$ tiene distribución t_{n-2} y, por esto, las inferencias acerca de $E(Y_k)$ se realizan con la distribución *t*.

Intervalo de confianza para $E(Y_k)$

Un IC de 1 - α para $E(Y_k)$ es: $\hat{Y}_k \pm t_{n-2;1-\alpha/2} \cdot s\left(\hat{Y}_k\right)$.

Ejemplo 1. Buscar un IC_{90} para $E(Y_k)$ para X_k = 55. Hallamos la estimación puntual $\hat{Y}_k$: $\hat{Y}_{55} = 10.0 + 2.0 \cdot (55) = 120$.

Luego, $s(\hat{Y}_k)$: $s^2\left(\hat{Y}_{55}\right) = 7.5 \cdot \left[\frac{1}{10} + \frac{(55-50)^2}{3400}\right] = 0.80515$,

de manera que

$s\left(\hat{Y}_{55}\right) = 0.89730$.

Para un coeficiente de confianza del 90% tenemos $t_{8;0.95}$ = 1.860. Luego, el IC_{90} es:

$$120 - 1.860 \cdot (0.89730) \leq E(Y_{55}) \leq 120 + 1.860 \cdot (0.89730),$$
$$\text{es decir, } 118.3 \leq E(Y_{55}) \leq 121.7.$$

ANEXO 1

Ecuaciones Normales

Las ecuaciones normales pueden ser derivadas mediante el cálculo. Para un conjunto de observaciones muestrales dado, (X_i, Y_i), la cantidad Q de la página 95 es una función de β_0 y β_1. Obtenemos:

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum\left(Y_i - \beta_0 - \beta_1 \cdot X_i\right) \\ \dfrac{\partial Q}{\partial \beta_1} = -2\sum X_i\left(Y_i - \beta_0 - \beta_1 \cdot X_i\right) \end{cases}$$

Posteriormente, igualamos estas derivadas a 0, utilizando b_0 y b_1 para denotar los valores particulares de β_0 y β_1, respectivamente, que minimizan a Q:

$$\begin{cases} -2 \cdot \sum\left(Y_i - b_0 - b_1 \cdot X_i\right) = 0 \\ -2 \cdot \sum X_i\left(Y_i - b_0 - b_1 \cdot X_i\right) = 0 \end{cases}$$

Simplificando, obtenemos:

$$\begin{cases} \sum_{i=1}^{n}\left(Y_i - b_0 - b_1 \cdot X_i\right) = 0 \\ \sum_{i=1}^{n} X_i\left(Y_i - b_0 - b_1 \cdot X_i\right) = 0 \end{cases}$$

Disociando la suma obtenemos:

$$\begin{cases} \sum Y_i - nb_0 - b_1\sum X_i = 0 \\ \sum X_i Y_i - b_0\sum X_i - b_1\sum X_i^2 = 0 \end{cases}$$

de las cuales, reordenando los términos, se obtienen las ecuaciones normales [9].

El cálculo las derivadas segundas mostraría que, con los estimadores de mínimos cuadrados b_0 y b_1, lo que se obtuvo es un **mínimo**.

$$\begin{cases} \sum Y_i = n \cdot b_0 + b_1 \cdot \sum X_i \\ \sum X_i Y_i = b_0 \cdot \sum X_i + b_1 \cdot \sum X_i^2 \end{cases}$$

Como dijimos, el objetivo del método de mínimos cuadrados es hallar estimaciones b_0 y b_1 para β_0 y β_1, respectivamente, para las cuales ***Q*** sea **mínima**. Después de las correspondientes manipulaciones algebraicas (que aquí no detallaremos), se llega a las siguientes dos expresiones para b_0 y b_1:

$$b_1 = \frac{\sum (X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \quad \text{y} \quad b_0 = \frac{1}{n} \cdot \left(\sum Y_i - b_1 \cdot \sum X_i\right) = \overline{Y} - b_1 \cdot \overline{X}$$

donde $\overline{X}$ e $\overline{Y}$ son las medias de X e Y, respectivamente.

Ejercicios.

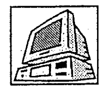
8.1 Para estudiar la asociación entre el consumo de sal y la presión arterial se seleccionaron 6 voluntarios entre los estudiantes de una Universidad, a cada uno se le administró una dosis determinada de sal en la dieta y se midió su presión arterial después de un tiempo de tratamiento. A continuación se presentan los datos obtenidos en el experimento y los resultados de un análisis de regresión lineal simple realizado con los mismos:

Sal [g/día]	Presión [mm Hg]
1,8	100
2,2	98
3,5	110
4,0	110
4,3	112
5,0	120

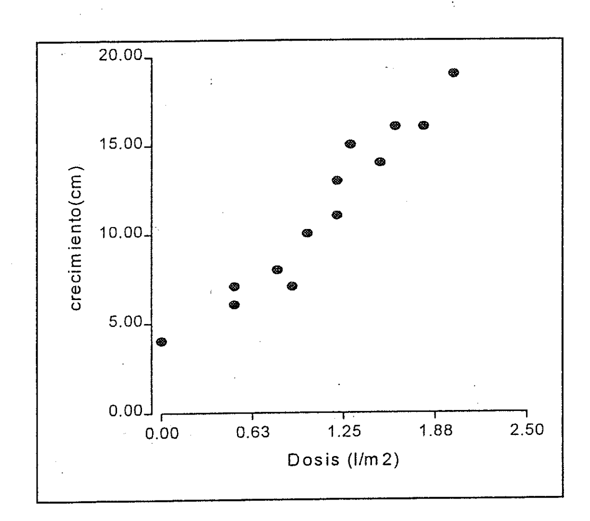
a. *Identificar la población y las unidades muestrales.*
b. *Identificar la variable independiente y la variable dependiente. Señalar cuál es aleatoria y cuál no y explicar por qué.*
c. *Construir el gráfico de dispersión e interpretarlo.*
d. *Escribir un modelo de regresión lineal apropiado para este estudio.*
e. *Explicar el significado de cada parámetro en términos del problema. Aclarar las unidades de cada parámetro..*
f. *Calcular los estimadores de mínimos cuadrados de los parámetros del modelo.*
g. *Estimar la varianza de la variable dependiente.*
h. *Construir intervalos del 95 % confianza para los parámetros del modelo.*
i. *Poner a prueba la hipótesis nula: No hay asociación entre el consumo de sal y la presión arterial.*
j. *Calcular e interpretar el coeficiente de determinación.*

k. *Calcular el residual correspondiente a la segunda observación.*

l. *Estimar la presión esperada para individuos que consumen 2,5 gr de sal por día con un intervalo del 95% de confianza*



8.2 Un productor hortícola necesita establecer el *crecimiento* esperado de una nueva variedad de repollo con las *dosis de fertilizante* de uso habitual en otras variedades de la misma especie. Para comprobarlo, realiza un experimento en 14 parcelas cultivadas con la nueva variedad. En cada parcela aplica una dosis determinada de fertilizante y mide el crecimiento promedio del diámetro de la hortaliza al cabo de tres semanas. A continuación se presenta el gráfico de dispersión y un cuadro con los datos obtenidos.



Parcela	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Dosis (l/m^2)	0.0	0.5	0.5	0.8	0.9	1.0	1.2	1.2	1.3	1.5	1.6	1.8	1.8	2.0
Crecimiento (cm)	4	7	6	8	7	10	11	13	15	14	16	16	16	19

a. *Identificar la población y las unidades muestrales.*

b. *Identificar a la variable independiente y la variable dependiente. Señalar cuál es aleatoria y cuál no y explicar por qué.*

c. *Proponer un modelo de regresión lineal simple para describir la relación estadística entre el crecimiento en diámetro de las plantas de repollo y la dosis de fertilizante.*

d. *Explicar por qué se trata de una relación estadística y no de una relación funcional.*

e. *Ajustar el modelo propuesto (esto es, estimar los parámetros β_0, β_1, y σ^2).*

f. *¿Que resultado da la prueba de la hipótesis $\beta_1 = 0$ con un nivel de significación $\alpha = 0{,}05$? Presentar la conclusión en términos de la interpretación biológica del problema.*

g. *Construir un intervalo del 95 % de confianza para crecimiento esperado de una planta de repollo que crece en una parcela tratada con 1,6 l/m^2 del fertilizante en cuestión.*

8.3 Con el fin de elaborar un modelo para predecir el rendimiento promedio de las plantaciones de una cepa de uva a partir del número de racimos promedio por planta al fin de la floración, se obtuvieron datos en 12

plantaciones de esta cepa tomadas al azar en el área de San Rafael (Mendoza). A continuación se presentan los datos obtenidos y los resultados de un análisis de regresión lineàl simple realizado a partir de los mismos.

Número de racimos por planta	116	83	111	97	116	80	125	116	117	93	107	122
Rendimiento [tn/ha]	5,6	3,2	4,5	4,2	5,2	2,7	4,8	4,9	4,7	4,1	4,4	5,4

Error típico	0,366
SCtotal	7,9825
SCerror	1,342

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción (b_0)	-1,02548909	0,78908011	-1,29960073	0,22289749
Num. de racimos (b_1)	0,05144651	0,00731375	7,03422011	3,5656E-05

a. *Identificar las unidades muestrales, la muestra y la población*
b. *Identificar la variable independiente y la variable respuesta (Notar que en este caso ambas variables son aleatorias. Por eso, la inferencia acerca de los parámetros del modelo es sólo aproximada).*
c. *Construir el gráfico de dispersión e interpretarlo.*
d. *Sobre el gráfico, dibujar la recta de regresión estimada.*
e. *Escribir el modelo de regresión lineal correspondiente al análisis presentado.*
f. *Aclarar las unidades en que se mide cada parámetro del modelo.*
g. *Construir un intervalo del 95% de confianza para la pendiente de la recta de regresión.*
h. *Explicar qué significa el intervalo construido.*
i. *Calcular e interpretar el coeficiente de determinación.*

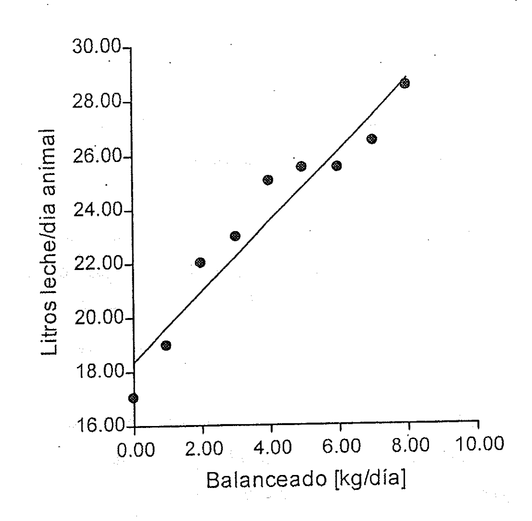
8.4 En una estación experimental de Rafaela, Santa Fe se realizó un ensayo para evaluar el efecto del nivel de *suplementación con alimento balanceado* (Kg/día/animal) sobre la *producción de leche* (lts/día/animal) de vacas de raza Holando-Argentina. Para ello se tomaron 9 vacas al azar dentro de un rodeo lechero, a cada una se asignó una dosis de alimento balanceado y se midió su productividad diaria promedio durante la lactancia. En las siguientes tablas y figuras se resumen los resultados obtenidos de un análisis de regresión lineal efectuado con los datos obtenidos

Analisis de Regresión Lineal

Variable	N	R^2
Leche	9	0.93

Coeficientes de regresión

Coef.	Estimad.	Error	$LI_{(95\%)}$	$LS_{(95\%)}$	*t*	*p*
Interc.	18.36	0.65	16.82	19.89	28.25	<0.0001
Pendiente	1.30	0.14	0.98	1.62	9.53	<0.0001



a. *Identificar las unidades muestrales, la población, la variable independiente y la variable aleatoria respuesta.*
b. *Escribir la ecuación de regresión lineal estimada e interpretar en términos agronómicos los estimadores de los parámetros. Indicar en el gráfico el valor de la ordenada al origen.*
c. *Según la ecuación propuesta en (a), ¿qué producción de leche promedio puede obtenerse con un nivel de suplemento de 5.5 Kg. diarios de balanceado por animal?*
d. *¿Cuál sería el valor esperado de la producción obtenida con una suplementación de 15kg de alimento balanceado por día? Comentar desde un punto de vista agronómico y estadístico su respuesta.*
e. *Interpretar el valor $p < 0.001$ asociado con la estimación de la pendiente.*

8.5 El nitrógeno es un nutriente fundamental para el crecimiento de las plantas porque forma parte de los pigmentos y enzimas que intervienen en la fotosíntesis. Un estudio de fisiología vegetal evaluó el contenido de Nitrógeno y el contenido de *clorofila,* el principal pigmento de la fotosíntesis, en 10 hojas de plantas de trigo seleccionadas al azar de diferentes macetas donde crecían con diferentes dosis de fertilizante nitrogenado. Ambas variables fueron medidas en milimoles /m² de hoja. A continuación se presentan los resultados de un análisis de regresión lineal simple realizado con los datos obtenidos

Estadísticos de la regresión	
R^2	0.87534459
Error típico	0.05761168
N	10

	Coeficientes	*EE*	*t*	*Valor p*	*LI(95%)*	*LS(95%)*
Constante	0,0472	0,0427	1,1050	0,3013	-0,0513	0.1458
Nitrógeno	0,0037	0,0005	3,4	≤ 0,0001	....	....

a. *Identificar las unidades muestrales, la muestra y la población.*
b. *Escribir el modelo de regresión y describir cada parámetro en términos del problema*

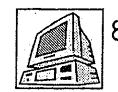
c. *Aclarar las unidades en que se mide cada parámetro del modelo.*
d. *Escribir la ecuación de regresión estimada.*
e. *Estimar la varianza de la variable dependiente. ¿Qué unidades tiene?*
f. *Calcular los valores que faltan en la tabla de resultados.*
g. *Formular las hipótesis necesarias para establecer si el contenido de clorofila de las hojas de trigo aumenta con su contenido de Nitrógeno en términos de los parámetros del modelo.*
h. *¿Que resultado da la prueba de estas hipótesis con un nivel de significación α=0,05? Justificar. (Notar que en este caso ambas variables son aleatorias. Por eso, la inferencia acerca de los parámetros del modelo es sólo aproximada).*

8.6 El exceso de fertilización nitrogenada puede provocar serios problemas ambientales. Cuando las plantas no alcanzan a absorberlo, parte del nitrógeno aplicado llegar al agua subterránea y contaminarla. Un estudiante investigó este tema en su trabajo de intensificación. Para ello, tomo una muestra aleatoria de 21 establecimientos del partido de Baradero y en cada uno registró la dosis promedio de fertilizante aplicada en los últimos 20 años (kg/ha/año) y el contenido actual de Nitratos (ppm) en el agua subterránea. Con los datos obtenidos, realizó un análisis de regresión lineal simple para establecer si el nivel de contaminación nitrogenada del agua subterránea depende del volumen de fertilizante aplicado. Los resultados del análisis figuran a continuación.

Estadísticos de la regresión	
Coeficiente de determinación R^2	0, 9652
Error Típico	11,7662
Nro. de observaciones	21

	Coeficientes	*Error Típico*
Intercepción	15,6643	4,6288
Dosis fertilizante	0,4129	0,0257

a. *Identificar la población, la muestra y las unidades muestrales*
b. *Escribir el modelo de regresión lineal correspondiente a este análisis y explicar el significado de cada parámetro en términos del problema.*
c. *Poner a prueba la hipótesis nula: No hay asociación entre el en contenido de nitratos del agua subterránea y la dosis de fertilizante promedio aplicada. Concluir en términos del problema (Notar que en este caso ambas variables son aleatorias. Por eso, la inferencia acerca de los parámetros del modelo es sólo aproximada).*
d. *Construir un intervalo del 95% de confianza para el contenido de nitratos del agua subterránea de un establecimiento que no aplica fertilizante.*



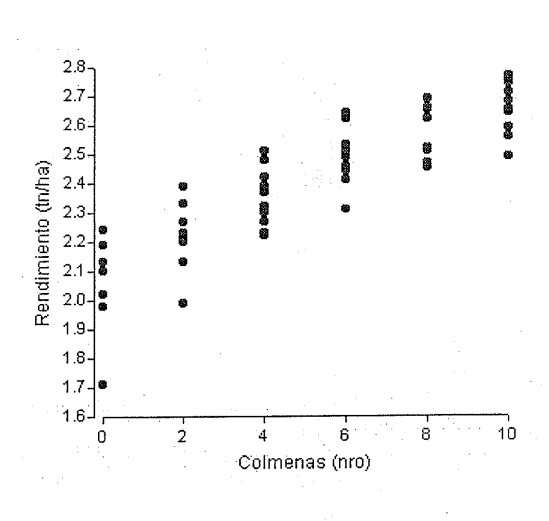
8.7 En el marco de un estudio sobre manejo de la fertilidad del suelo, se realizó un experimento para evaluar los efectos de la aplicación de fertilizantes orgánicos dentro de un campo experimental. Para ello, se delimitaron 16 parcelas 400m² cada una y a cada una se le asignó al azar una dosis de de compost de residuos urbanos Las dosis aplicadas fueron 0, 6, 12 y 36 tn/ha. En el siguiente cuadro se indican las cantidades de nitrógeno inorgánico (Kg/ha de Nitratos + Amonio) en los primeros 20cm del suelo medidas en cada parcela 1 año después del tratamiento:

Parcela	Dosis compost (Tn/ha)	Nitrógeno Inorgánico (Kg/ha)
1	0	180
2	0	153
3	0	152
4	0	140
5	6	195
6	6	185
7	6	150
8	6	175
9	12	195
10	12	165
11	12	200
12	12	175
13	18	188
14	18	214
15	18	204
16	18	199

a. *Dibujar un esquema de como pudo haber estado distribuido el experimento en el campo. Discutir la forma en que cada dosis de compost fue asignada a cada una parcela.*
b. *Realizar un gráfico de dispersión y comentarlo.*
c. *Escribir el modelo lineal correspondiente y describir cada parámetro en términos del problema. Calcular y graficar la recta de regresión.*
d. *Estimar la varianza de la variable dependiente. ¿Qué unidades tiene?*
e. *Construir intervalos del 95 % de confianza para los parámetros. ¿Que unidades tienen sus extremos?*
f. *Poner a prueba las hipótesis de que los parámetros del modelo valen cero (α = 0.05). ¿Cómo se interpreta cada hipótesis?*
g. *Para una dosis de 3 Tn/Ha de compost, ¿cuál sería el total de nitrógeno en el suelo? ¿Y para una dosis de 25 Tn/Ha?*
h. *Calcular e interpretar el residual para la 5ª observación.*
i. *Calcular e interpretar el coeficiente de determinación.*

8.8 El girasol es una planta de polinización entomófila, esto significa que, para que produzca semillas, sus flores deben ser visitadas por insectos que transportan el polen. Por este motivo, el rendimiento de los cultivos de girasol depende críticamente de la actividad de los insectos

polinizadores. Con frecuencia, los insectos que se encuentran naturalmente en los lotes cultivados no alcanzan a polinizar todas las flores y por eso el rendimiento aumenta si en ellos se instalan colmenas de abejas. En una cooperativa agrícola, los productores condujeron un estudio para evaluar la relación entre densidad de abejas y el rendimiento de sus cultivos de girasol. Para ello, seleccionaron al azar 6 grupos de 10 lotes sembrados y en los lotes de cada grupo instalaron respectivamente 0, 2, 4, 6, 8 y 10 colmenas por ha. Luego registraron el rendimiento obtenido en cada lote y se realizaron un análisis de regresión lineal simple. A continuación se presenta un gráfico de dispersión con los datos obtenidos y parte de los resultados del análisis:



	Estimadores	*Error Típico*
b_0 =	_____	0.025
b_1 =	0.058	0.004

$$\sum(Y_i - \bar{Y})^2 = 3{,}062$$

$$\sum\left(Y_i - (b_0 + b_1 X_i)\right)^2 = 0{,}711$$

$$\sum(X_i - \bar{X})^2 = 700$$

$$\bar{X} = 5$$

$$\bar{Y} = 2{,}396$$

N = 60

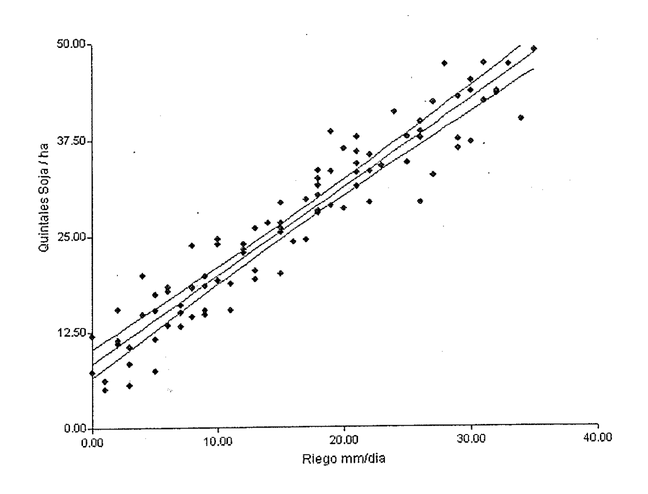
a. *¿Cuál fue el objetivo del ensayo?*
b. *Identificar a las poblaciones, las muestras, la variable aleatoria consideradas en este problema*
c. *Escribir el modelo de regresión lineal simple y explicar el significado de cada parámetro en términos del problema.*
d. *¿Cuál es el dominio del modelo?*
e. *Agregar en el gráfico la recta de regresión estimada.*
f. *Estimar el residual no explicado por el modelo para un lote que tenía 8 colmenas/ha y rindió 2.41 tn/ha.*
g. *Estimar la varianza de la variable aleatoria.*
h. *Poner a prueba la hipótesis nula: No hay asociación entre el rendimiento del girasol y el número de colmenas/ha.*
i. *Calcular en coeficiente de determinación e interpretarlo en términos del problema.*

8.9 Los técnicos de una empresa evalúan alternativas para el alquiler de un equipo para regar 100 has de cultivo de soja. Las opciones disponibles son 3 equipos con diferente capacidad (mm/día) y diferente costo de uso total en toda la campaña ($): Equipo A. (Capacidad = 10 mm/día, Costo = $ 15000.-), Equipo B. (Capacidad = 20 mm/día, Costo = $ 20000.-) y Equipo C. (Capacidad = 30 mm/día, Costo = $ 35000.-). Los técnicos saben que el precio neto de venta de la soja es de $12/quintal y cuentan con información de experiencias realizadas en la misma localidad que permitieron estimar la asociación entre el rendimiento de

la soja (quintales/ha) y la intensidad de riego (mm/día). A continuación se presentan los principales resultados del análisis de dicha información:

Coeficientes de regresión y estadísticos asociados

Coef	Est.	EE	LI(95%)	LS(95%)	T	p-valor
const	8.39	0.72	6.97	9.81	11.71	<0.0001
Riego(mm/dia)	1.15	0.04	1.07	1.22	30.10	<0.0001



a. *¿La información disponible es suficiente para aceptar que el rendimiento esperado del cultivo de soja aumenta con la intensidad de riego aplicada? Justificar la respuesta*

b. *Producir intervalos del 95 % de confianza para el rendimiento esperado del cultivo de soja con cada uno de los tres equipos de riego y de un cultivo no regado.*

c. *¿Cuál equipo de riego elegiría? Justificar la respuesta*

d. *¿Qué utilidad tienen los datos disponibles para estimar el rendimiento de soja regada con 100 mm/día? Justificar la respuesta*

e. *En caso de que el precio de la soja aumentara, ¿cuál debería ser este aumento para que usted cambie de opinión sobre el equipo de riego a elegir?*

f. *¿Cuál debería ser el aumento del precio neto de la soja para que no conviniera alquilar ningún equipo de riego?*

ANALISIS DE DATOS CATEGORICOS

La distribución χ^2 tiene un gran campo de aplicación en el análisis de variables de naturaleza categórica, es decir, cuando se trata de datos de **frecuencia**. En ciencia e ingeniería, muchas veces se cuenta con información acerca de la cantidad de veces que aparece una determinada característica en una muestra y en esta clase se verá cómo se puede recurrir al empleo de la distribución χ^2 para analizar este tipo de datos. Concretamente, se verán dos aplicaciones directas: (i) las pruebas de **bondad del ajuste**, y (ii) **tablas de contingencia**. Entre estas últimas veremos las pruebas de **homogeneidad** y las pruebas de **independencia**.

Pruebas de Bondad del Ajuste

Estas pruebas se aplican cuando se desea contrastar una distribución de frecuencias **observada** en una muestra con una distribución de frecuencias **teórica** o que responde a un determinado modelo o situación preconcebida. Para aplicar la prueba de χ^2 de bondad del ajuste se necesita una tabla donde se encuentren registradas las frecuencias observadas y las frecuencias teóricas o esperadas según el modelo. El estadístico que se utiliza en estas pruebas es el siguiente:

$$\chi_v^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \quad (9.1)$$

donde k es el número de categorías y o_i y e_i son las frecuencia observada y esperada en la i-ésima categoría, respectivamente. Este estadístico tiene una distribución χ^2 con un número de grados de libertad (v) igual a la cantidad de categorías menos 1. Una aclaración muy importante: **tanto o_i como e_i deben ser frecuencias absolutas**, no frecuencias relativas o proporciones.

Ejemplo.
Son conocidos en Genética los experimentos clásicos conducidos por Mendel en los albores de esa ciencia, en los que se buscaba determinar el modo de herencia de una serie de caracteres cualitativos observados en plantas de arveja. Uno de los caracteres estudiados por Mendel era el tipo de tegumento de la semilla. Mendel tenía arvejas con dos tipos de tegumento: rugoso y liso. Según su hipótesis, en cruzamientos realizados entre ciertos tipos de plantas, el esperaba que aparecieran en la descendencia de dichos cruzamientos, arvejas de tegumento liso y rugoso en la proporción 3:1, es decir, 3 semillas de tegumento liso por cada semilla de tegumento rugoso. Supongamos que en un experimento en el cual se obtiene una descendencia compuesta por 100 semillas, un genetista encuentra 285 semillas de tegumento liso y 115 de tegumento rugoso. ¿Sería razonable, con α = 0.05, pensar que esa proporción observada no está demasiado alejada de la proporción 3:1 dictada por la ley de Mendel?

1. **Hipótesis**. H_0: la proporción es 3:1; H_1: la proporción no es 3:1.
2. **Nivel de significación**. α = 0.05.
3. **Estadística de la prueba**. $\chi_v^2 = \sum_{i=1}^{2} \frac{(o_i - e_i)^2}{e_i}$ que se

   distribuye como χ_1^2 puesto que, para esta prueba k = 2 y, por consiguiente, v = 2 – 1 = 1.

4. **Regla de decisión**. $P(\chi_1^2 > 3.84) = 0.05$. Rechazamos H_0 sí, y sólo sí, el valor de χ^2 calculado es mayor que 3.84. En caso contrario, se acepta H_0.

5. **Cálculos.**

Tabla 9.1.

Tegumento	o_i	e_i	$o_i - e_i$	$(o_i - e_i)^2/e_i$
Liso	285	400·(3/4) = 300	-15	0.75
Rugoso	115	400·(1/4) = 100	15	2.25
Total	400	400	---	**3.00**

6. **Decisión**. Puesto que 3.0 < 3.84 no puede rechazarse H_0 con α = 0.05. Los datos de la muestra no constituyen una prueba suficiente como para dudar de que las proporciones verdaderas son 3:1.

Tablas de contingencia

En una *tabla de contingencia* la información también está formada por cuentas o frecuencias organizadas en *f* filas y *c* columnas y se dice entonces que se tienen **dos criterios de clasificación.** Se pueden describir dos situaciones posibles.

(1) Hay ***f* poblaciones** de interés, cada una en una fila de la tabla, y en cada población se describen *c* **categorías** o **atributos**. Se toma una muestra de cada población y las frecuencias se anotan en las celdas de la tabla.

(2) Hay **una sola población** de interés y cada individuo es clasificado respecto a **dos factores** diferentes. Hay *f* categorías de un factor y *c* categorías del otro factor. Se toma una sola muestra y se anota el número de individuos en cada categoría de ambos factores.

Las situaciones de tipo (1) se conocen como **pruebas de homogeneidad** y las situaciones de tipo (2) como **pruebas de independencia**. Estas pruebas son muy parecidas; de hecho en ambas se utilizan tablas de contingencia y se calculan los valores esperados y los grados de libertad de manera similar. Lo que diferencia ambas pruebas son las **hipótesis**. El estadístico que se utiliza es el mismo que el empleado en las pruebas de bondad del ajuste:

$$\chi_\nu^2 = \sum_{j=1}^{c} \sum_{i=1}^{f} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{9.2}$$

donde *f* es el número de filas, *c* número de columnas, o_{ij} y e_{ij} son las frecuencia observada y esperada en la celda *ij*, respectivamente. Este estadístico tiene una distribución χ^2 con un número de grados de libertad igual a $\nu = (f - 1)\cdot(c - 1)$. Por ejemplo, si la tabla de contingencia fuera 2 × 2, tendríamos una cantidad de grados de libertad igual a $\nu = (2 - 1)\cdot(2 - 1) = 1$.

Pruebas de homogeneidad

Estas pruebas se utilizan cuando se desea determinar si las proporciones de las diferentes categorías son las mismas para todas las poblaciones. La hipótesis nula establece que las poblaciones son homogéneas con respecto a las categorías y la alternativa establece que no lo son. Otra manera de abordar el mismo problema es preguntar si las muestras provienen o no de la misma población.

Obtención de los valores esperados

Con base en la hipótesis nula, se espera que las proporciones de las distintas categorías dentro de cada población, son iguales para todas las poblaciones y, por tanto, a las proporciones marginales. Esto equivale a decir que para la celda *i,j* el número esperado será igual a:

$$e_{ij} = \frac{n_{i.}}{n_{..}} \cdot n_{.j} = \frac{n_{.j}}{n_{..}} \cdot n_{i.} \qquad (9.3)$$

donde $n_{i.}$ es el total de la fila *i*, $n_{.j}$ es el total de la columna *j*, y $n_{..}$ es el total general.

Ejemplo.
En la siguiente tabla se resume la información sobre el tipo de marcas encontradas en hojas de tréboles blancos muestreados en un sitio no pastoreado y en otro pastoreado. En cada sitio se muestrearon 550 y 450 individuos respectivamente.

Tabla 9.2.

		Tipo de marca				
		L	LL	Y	O	Total
Sitio	No pastoreado	409	11	22	8	**450**
	Pastoreado	512	4	14	20	**550**
	Total	**921**	**15**	**36**	**28**	**1000**

Viendo la forma en que es planteado el problema, una hipótesis nula apropiada que puede ponerse a prueba sería que la proporción de individuos con los diferentes tipos de marcas en las hojas es la misma para las dos poblaciones, o sea en cada sitio.

Luego:

1. **Hipótesis**: H_0: $p_{1j} = p_{2j}$; donde j = 1,2,3,4 son las 4 marcas e i = 1,2 son los dos sitios.
   H_1: $p_{1j} \neq p_{2j}$
2. **Nivel de significación**. $\alpha = 0.05$.
3. **Estadística de la prueba.** $\chi_v^2 = \sum_{i=1}^{f\times c} \frac{(o_i - e_i)^2}{e_i}$ que se distribuye aproximadamente como χ_3^2. Aquí $v = (2-1)\cdot(4-1) = 3$.
4. **Regla de decisión**. $P(\chi_3^2 > 7.81) = 0.05$. Rechazamos H_0 si, y solo si, el valor de χ^2 calculado es mayor que 7.81. En caso contrario, se acepta H_0.
5. **Cálculos.**

$$\chi_3^2 = \sum_{i=1}^{f\times c} \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(409-414.45)^2}{414.45} + \frac{(11-6.75)^2}{6.75} + \cdots + \frac{(20-15.4)^2}{15.4}$$

$$\cong 11.82$$

6. **Decisión**. Puesto que 11.82 > 7.81 se rechaza H_0 con $\alpha = 0.05$. La proporción de individuos con diferentes tipos de marcas no es la misma en las dos poblaciones o sea que las dos poblaciones de tréboles no son homogéneas en cuanto a su distribución de marcas.

Pruebas de independencia

Este tipo de prueba se aplica cuando existe interés en determinar si dos atributos categóricos presentan algún tipo de asociación entre ellos o, si por el contrario, son independientes. En otras palabras concentramos nuestra atención en la relación entre dos factores diferentes de la misma población. En esta prueba tomamos una muestra de la población y caracterizamos cada individuo según dos criterios de clasificación dispuestos en *i* filas y *j* columnas. A diferencia de las pruebas de homogeneidad donde en muchos casos los totales de filas están fijos por anticipado, en las pruebas de independencia solo el tamaño muestral es fijo y tanto los totales de filas como los de columnas son variables aleatorias. La hipótesis nula establece que la categoría de un individuo con respecto al factor *A* es independiente de la categoría con respecto al factor *B*. En otras palabras y recordando el capítulo de probabilidades, la hipótesis nula establece que los eventos son independientes y por lo tanto $P(A \cap B) = P(A) \cdot P(B)$.

Ejemplo.
En el partido de Balcarce se realizó una encuesta a 930 productores de trigo-soja y se los clasificó según el método de siembra empleado (siembra convencional o siembra directa) y el área sembrada. Se consideraron 3 categorías: (1) área menor a 100 ha; (2) área entre 100 y 500 ha y;(3) área superior a 1000 ha. Los resultados se muestran en la siguiente tabla de contingencia:

Tabla 9.3.

		Tipo de siembra		
		SC	SD	Total
área	1	94	180	**274**
	2	116	320	**436**
	3	140	80	**220**
	Total	**350**	**580**	**930**

Si el método de siembra y el área sembrada son independientes, esperaríamos que la proporción de productores que usan siembra convencional sea (350/930) = 0.376, sea cual fuere el área sembrada. Y, por ejemplo, el número esperado productores que usan siembra convencional y tienen un área sembrada reducida (categoría 1) sería: (274)·(350/930) = 103.1. Las frecuencias esperadas para nuestro ejemplo entonces son:

Tabla 9.4.

		Tipo de siembra		
		SC	SD	Total
área	1	103.12	170.88	**274**
	2	164.09	271.91	**436**
	3	82.8	137.20	**220**
	Total	**350**	**580**	**930**

y dado que la tabla de contingencia es una tabla a 3 × 2, tenemos 2 grados de libertad. Con esta evidencia obtenida en la muestra, ¿se puede sostener la hipótesis de que el *método de siembra* y el *área sembrada* son independientes ($\alpha = 0.01$)?

1. **Hipótesis**. H_0: el *método de siembra* y el *área sembrada* son independientes. H_1: están relacionados (son dependientes). H_0: $p_{ij} = p_{i.} \cdot p_{.j}\ \forall\, i, \forall\, j$; H_1: $p_{ij} \neq p_{i.} \cdot p_{.j}$ para algún par i,j.

2. **Nivel de significación**. $\alpha = 0.01$.

3. **Estadística de la prueba.** $\chi_v^2 = \sum_{i=1}^{f \times c} \frac{(o_i - e_i)^2}{e_i}$ que se distribuye como χ_2^2.

4. **Regla de decisión**. $P(\chi_2^2 > 9.21) = 0.01$. Rechazamos H_0 si, y solo si, el valor de χ^2 calculado es mayor que 9.21. En caso contrario, se acepta H_0.

5. **Cálculos.**

$$\chi_2^2 = \sum_{i=1}^{f \times c} \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(94 - 113.12)^2}{113.12} + \frac{(180 - 170.88)^2}{170.88} + \ldots\ldots + \frac{(80 - 137.20)^2}{137.20}$$

$$= 87.26$$

1 **Decisión**. Dado que 87.26 > 9.21 se rechaza H_0 con $\alpha = 0.01$. Hay evidencia suficiente para rechazar la hipótesis de que el método de labranza y el área sembrada son independientes.

Ejercicios

9.1 Alber's fabrica y distribuye tres tipos de cerveza: Lager, Pilsen y Stout. En un análisis de segmentación de mercado para las tres cervezas, el grupo de investigación encargado ha planteado la duda de si las preferencias para las tres cervezas son diferentes entre los consumidores hombres y mujeres. Si la elección del tipo de cerveza fuera independiente del género del consumidor, se iniciaría una campaña de publicidad para todas las cervezas de Alber's. Sin embargo, si la elección depende del género del consumidor, se ajustarán las promociones para tener en cuenta los distintos mercados meta. Se toma una muestra aleatoria de 150 bebedores de cerveza y después de saborear cada una, se les pide expresar su preferencia o primera alternativa. En base a estos resultados, realizar un breve informe para presentar en el departamento de publicidad:

		Cerveza preferida		
		Lager	Pilsen	Stout
Género	Masculino	20	40	20
	Femenino	30	30	10

$H_0)\ P(i \cap j) = P_i \cdot P_j$

80
70

50 70 30

9.2 Un semillero intenta probar un híbrido nuevo de maíz aparentemente resistente a heladas. Para ello se escogen 279 parcelas donde se realiza una siembra temprana (alta probabilidad de heladas). 139 parcelas escogidas al azar son sembradas con el híbrido tradicional y las otras 140 son sembradas con el híbrido nuevo. Luego de transcurrido el período de heladas, se comprobó que en 31 parcelas sembradas con el híbrido tradicional se observaron problemas de densidad de cultivo (debido a muerte de plantas por heladas), mientras que 17 parcelas sembradas con el híbrido nuevo presentaron este problema. Concluir con respecto a la resistencia a heladas de estos dos híbridos. Si un productor quiere sembrar temprano el maíz y le consulta sobre cuál híbrido utilizar (tradicional o nuevo), ¿qué le sugeriría? ¿En qué basaría su respuesta?

9.3 Otro estudio dirigido a comparar los resultados de los métodos de labranza química y mecánica examinó su asociación con la cantidad de malezas presentes en los cultivos de maíz. Para ello, se tomó una muestra aleatoria de 100 lotes de maíz en el partido de Pergamino y se los clasificó según el tipo de labranza (química o mecánica) y el grado de infestación con malezas (alto, medio o bajo). Los datos obtenidos son presentados en la tabla:

		Grado de infestación con malezas		
		Alto	Medio	Bajo
Tipo de Labranza	Química	19	24	17
	Mecánica	8	14	18

a. Identificar a las unidades muestrales, la población y la muestra.
b. ¿A qué tipo de prueba corresponde la hipótesis nula: no hay asociación entre el tipo de labranza y el grado de infestación con malezas?
c. Poner a prueba esta hipótesis y concluir en términos del problema.
d. Explicar qué es tipo de error se puede haber cometido en este análisis

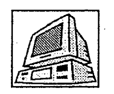
9.4 Un nuevo producto funguicida es promocionado asegurando que aumenta la tolerancia de las plantas de trigo a cierto hongo. Para evaluar este fungicida, los técnicos del INTA tomaron 100 plantas de trigo y las infectaron con el hongo. Luego, trataron con el funguicida 50 de estas plantas seleccionadas al azar y un tiempo después evaluaron el estado de las plantas. Entre las 50 plantas tratadas con funguicidas, 10 se encontraban en buen estado, 20 se encontraban levemente afectadas y 20 se encontraban en muy mal estado. Entre las plantas que no recibieron funguicida, 5 se encontraban en buen estado, 20 se encontraban levemente afectadas y 25 se encontraban en muy mal estado.

b) $P_{1j} = P_{2j}$

a. Estimar las proporciones esperadas de plantas que mantienen buen estado con y sin tratamiento con el fungicida.
b. Explicar por qué existe incertidumbre en las estimaciones anteriores.

c. *Realizar un análisis apropiado para determinar si los resultados de esta experiencia demuestran que el funguicida modifica la tolerancia de las plantas al hongo.*
d. *Comunicar en lenguaje coloquial la conclusión a la que arriban los técnicos*
e. *Explicar qué tipo de error podrían haber cometido.*

9.5 Antes de autorizar la exportación de arándanos a los Estados Unidos, se realizó un control de calidad para evaluar la calidad de los cargamentos provenientes de diferentes provincias. Para ello se obtuvieron muestras aleatorias de cargamentos provenientes de Tucumán, Buenos Aires y Santa Fé y se determinó cuantos de ellos cumplían con las normas de calidad fijadas con respecto al tamaño de las frutas. En la siguiente tabla se presentan los datos obtenidos sobre el número de cargamentos que superaban o no el tamaño mínimo provenientes de tres provincias productoras.



Número de cargamentos

	Procedencia		
	Tucumán	Buenos Aires	Santa Fé
Superior al tamaño mínimo	22	33	32
Inferior al tamaño mínimo	20	10	8

a. *Identificar las poblaciones, las muestras y las unidades de observación.*
b. *¿Cual es la variable aleatoria analizada? ¿De qué tipo de variable se trata?*
c. *Estimar las proporciones de cargamentos de cada provincia que no superan el tamaño de las frutas mínimo para exportación.*
d. *¿Por qué existe incertidumbre respecto de estas estimaciones?*
e. *Formular y poner a prueba las hipótesis necesarias para inferir si las proporciones de cargamentos que no superan el tamaño mínimo de las frutas difiere entre provincias (a = 0,05).*
f. *¿Cuál provincia tuvo más cargamentos con frutas menores al tamaño mínimo que lo esperado bajo la hipótesis nula?*

9.6 La siguiente tabla muestra los datos de un estudio medico en le cual se tomaron al azar 119 partos ocurridos en 2005 en la ciudad de Buenos Aires y se registró si las madres eran o no fumadoras y si su bebé tenía peso normal o peso bajo (menor que de 2.5 kg).

Número de partos

		Hijo	
		Peso Normal	Peso Bajo
Madre	Fumadora	3	13
	No Fumadora	57	46

Denominemos "A" al evento "madre fumadora" y "B" al evento "bebé nacido con peso normal".

a. *Identificar las unidades muestrales, la muestra y la población.*
b. *Estimar P[B / A] y P[B / AC].*
c. *Formular y poner a prueba la hipótesis nula: La ocurrencia de bajo peso al nacer de sus bebes es independiente del hábito de fumar de las madres.*
d. *¿Qué tipo de error se puede haber cometido en la prueba de hipótesis anterior? Explicar su significado en términos del problema.*

EJERCICIOS ADICIONALES CON ALGUNAS RESPUESTAS

[1] En un campo se pesaron 11 novillos. Al final de la operación se obtuvieron los siguientes datos:

Individuo	Peso [Kg/animal]
12453	450
13458	375
854	350
1234	425
5864	400
84952	415
12448	380
13221	395
953	430
7531	440
1035	390

El criterio establecido por la agroempresa para enviar la hacienda al remate/feria es un peso mayor a 390 Kg. Responda:

a. ¿Qué cantidad de animales de este lote será vendida?
   7 animales.
b. ¿Qué porcentaje representa?
   63.63%
c. Construya un histograma de frecuencias a partir de los datos de la tabla con solo dos clases (*0-390 Kg*, y *más de 390 Kg*).
d. Construya un diagrama de caja y bigotes para todos los datos.

[2] Se realizó una encuesta a 30 productores rurales de la zona norte de Buenos Aires y Sur de Santa Fe. A cada productor se le preguntó qué tipo de producción tenía (agrícola, ganadera o mixta). A partir de los resultados genere una tabla de doble entrada y conteste:

PRODUCTOR	PROVINCIA	TIPO
1	BUENOS AIRES	MIXTO
2	BUENOS AIRES	MIXTO
3	BUENOS AIRES	MIXTO
4	BUENOS AIRES	AGRICOLA
5	BUENOS AIRES	AGRICOLA
6	BUENOS AIRES	MIXTO
7	BUENOS AIRES	MIXTO
8	BUENOS AIRES	GANADERO
9	BUENOS AIRES	MIXTO
10	BUENOS AIRES	GANADERO
11	BUENOS AIRES	MIXTO
12	BUENOS AIRES	MIXTO
13	BUENOS AIRES	AGRICOLA
14	BUENOS AIRES	MIXTO
15	BUENOS AIRES	AGRICOLA
16	SANTA FE	MIXTO
17	SANTA FE	MIXTO
18	SANTA FE	AGRÍCOLA
19	SANTA FE	MIXTO
20	SANTA FE	MIXTO
21	SANTA FE	MIXTO
22	SANTA FE	MIXTO
23	SANTA FE	MIXTO
24	SANTA FE	MIXTO
25	SANTA FE	MIXTO
26	SANTA FE	AGRÍCOLA
27	SANTA FE	MIXTO
28	SANTA FE	AGRÍCOLA
29	SANTA FE	MIXTO
30	SANTA FE	AGRÍCOLA

a. ¿Qué porcentaje del total de productores tiene una explotación mixta?
   66,67%
b. De los productores con explotación netamente agrícola, ¿qué porcentaje se encuentra en la provincia de Buenos Aires?
   50%.

[3] Dos candidatos a los consejos de administración *A* y *B*, compiten por el control de una corporación. Las probabilidades de ganar de estos candidatos son 0.7 y 0.3, respectivamente. Si gana *A*, la probabilidad de introducir un nuevo producto es 0.8; si gana *B*, la correspondiente probabilidad es 0.4. Demuestre que, antes de las elecciones, la probabilidad de que sea introducido un nuevo producto es igual a 0.68.

[4] Un productor desea presentarse a una licitación de granos embolsados y por ello presta especial atención a que el peso de cada bolsa no se aparte excesivamente del promedio. Si el promedio es de 63 kg con un desvío estándar de 2 kg:

a. ¿Cuál es la probabilidad de que una bolsa no se aparte más de 3 kg. del promedio?
   0.8664
b. Si se toma una bolsa al azar, ¿cuál es la probabilidad de que pese menos de 60 kg.?
   0.0668
c. Si se toma al azar un lote de 10 bolsas, ¿cuál es la probabilidad de que a lo sumo una pese menos de 60 kg.?
   0.85945
d. Si se toma un lote de 1000 bolsas, ¿cuál es la probabilidad de que a lo sumo 100 pesen menos de 60 kg.?
   Aproximadamente 1

[5] Sea *X* una variable aleatoria con distribución binomial, con $n = 10$ y $\pi = 0.5$.

a. Determinar las probabilidades de que *X* se encuentre dentro de una desviación estándar de la media y a dos desviaciones estándares de la media.
   x: 4 a 6, $p = 0.65625$
b. ¿Cómo cambiarían las respuestas en (a) si $n = 15$ y $\pi = 0.4$?
   x: 2 a 8, $p = 0.9785$

[6] Sea *Z* una variable aleatoria normal estándar. Hallar:

a. $P(Z < 1.20)$; **0.8849**
b. $P(Z > 1.33)$; **0.0918**
c. $P(Z < -1.70)$; **0.0446**
d. $P(Z > -1.00)$; **0.8413**
e. $P(1.20 < Z < 1.33)$; **0.0233**
f. $P(-1.70 < Z < 1.20)$; **0.8403**
g. $P(-1.70 < Z < -1.00)$ **0.1141**

[7] Una compañía recibe un lote de insumos muy grande. Se analiza una muestra aleatoria de 16 artículos y se acepta el lote si menos de dos resultan defectuosos. ¿Cuál es la probabilidad de aceptar un envío que contenga:

a. 5% de artículos defectuosos? **0.8107**
b. 15% de artículos defectuosos? **0.2840**
c. 25% de artículos defectuosos? **0.0635**

[8] Se sabe que el dinero que gastan al año los estudiantes de la Facultad de Agronomía en libros de texto sigue una distribución normal con media \$ 125 y desviación típica \$ 25.

a. ¿Cuál es la probabilidad de que un estudiante elegido al azar gaste menos de \$ 60 en libros de texto al año?
   0.0047

b. ¿Cuál es la probabilidad de que un estudiante elegido al azar gaste más de $ 150 en libros de texto al año?
   0.1587
c. ¿Cuál es la probabilidad de que un estudiante elegido al azar gaste entre $ 80 y $ 135 en libros de texto al año?
   0.6195
d. Se quiere encontrar un rango de gastos en libros en el cual se incluyan el 80% de los estudiantes de esta universidad. Explicar por qué pueden encontrarse infinitos rangos que cumplan esta condición, y encontrar el rango menos costoso.
   el rango menos costoso: de $0 a $146.04

[9] El 90% de los árboles plantados en un monte forestal sobreviven hasta la tala final del rodal. ¿Cuál es la probabilidad de que sobrevivan 10 o más entre 15 árboles que acaban de ser plantados?
p = **0.99776**

[10] La duración de lámparas de luz producidas por un cierto fabricante tiene una media de mil doscientas horas y una desviación típica de cuatrocientas horas y se sabe que la población sigue una distribución normal. Supongamos que adquirimos 9 lámparas, que pueden ser consideradas como una muestra aleatoria de la producción del fabricante.

a. ¿Cuál es la esperanza de la media muestral de la duración de estas lámparas?
   1200
b. ¿Cuál es la varianza de la media muestral?
   17777
c. ¿Cuál es el error estándar de la media muestral?
   133.33
d. ¿Cuál es la probabilidad de que el tiempo medio de duración de las lámparas adquiridas sea menor a 1050 horas?
   0.13

[11] El dueño de una tienda de discos ha comprobado que el 20% de los clientes que entran en su tienda realizan una compra. Cierta mañana, entraron en esta tienda 180 personas, que pueden ser consideradas como una muestra aleatoria de todos sus clientes.

a. ¿Cuál será la media de la proporción muestral de clientes que realizaron alguna compra?
   0.20
b. ¿Cuál es la varianza de la proporción muestral?
   0.000889
c. ¿Cuál es el error estándar de la proporción muestral?
   0.0298
d. ¿Cuál es la probabilidad de que la proporción muestral sea menor que 0,15?)
   0.04669

[12] Una corporación ha recibido 120 solicitudes de trabajo de estudiantes que acaban de terminar su carrera de agronomía. Suponiendo que estas solicitudes pueden ser consideradas como una muestra aleatoria de todos los ingenieros, ¿cuál es la probabilidad de que entre un 35% y un 45% de las solicitudes correspondan a mujeres si se sabe que el 40% de los ingenieros agrónomos que acaban de terminar su carrera son mujeres?
0.8686

[13] Suponga que una muestra aleatoria de tamaño $n = 25$, es seleccionada de una población con media μ, y desvío standard σ. Para los siguientes valores de μ y σ, determine los valores de $\mu_{\bar{x}}$ y $\sigma_{\bar{x}}$.

a. $\mu = 100$ y $\sigma = 50$;

$\mu_{\bar{x}} = 100$ y $\sigma_{\bar{x}} = 10$

b. $\mu = 750$ y $\sigma = 25$.

$\mu_{\bar{x}} = 750$ y $\sigma_{\bar{x}} = 5$.

[14] Después de seleccionar una muestra y calcular el IC para μ, una persona dice: "tengo una confianza del 88% de que la media de la población fluctúa entre 106 y 122". ¿Qué es lo que realmente está diciendo?

a. ¿que hay una probabilidad de 0.88 de que μ fluctúe ente 106 y 122?
b. ¿qué hay una probabilidad de 0.88 de que el valor real de μ sea 114 (el punto medio del intervalo)?
c. ¿qué el 88% de los intervalos obtenidos de las muestras de este tamaño contendrán la media de la población?
d. (a), (b) y (c) son correctas.

Rta: c

[15] Una muestra de 30 parcelas sembradas con algodón, arrojó un rendimiento medio de 950 kg/ha. Sabiendo que los rendimientos tienen distribución normal con desvío típico 25 kg/ha, estimar el verdadero rendimiento medio, mediante un IC_{95} y un IC_{99}.

IC_{95}: 950 ± 8.95kg ; IC_{99}: 950 ± 11.73kg

[16] Un ensayo de un nuevo híbrido de maíz arrojó los siguientes resultados (Tn/ha).

12.4	11.0	10.5	11.7	9.9	12.0	8.9	9.7	11.5	11.1

¿Estos resultados constituyen evidencia suficiente para afirmar que este híbrido es mejor que otro que tiene una media de rendimiento de 10 Tn/ha? ($\alpha = 0.05$).

Prueba de hipótesis de una cola, varianza estimada a partir de la varianza muestral. Valor t = 2.48, valor p = 0.0176. Los resultados aportan evidencia suficiente para afirmar que los rindes del nuevo híbrido son mayores a 10 Tn/ha.

[17] Se estudian dos raciones, A y B, para el engorde de cerdos. Se tomaron 8 lotes de cerdos, cada uno formado por hermanos de la misma lechigada, y se le asignaron las raciones aleatoriamente en cada lote. Los resultados, en kg, se presentan en la siguiente tabla:

		Lotes							
		1	2	3	4	5	6	7	8
Raciones	A	75	80	80	72	72	75	78	82
	B	85	79	90	68	75	81	88	90

a. Probar si ambas raciones producen igual engorde ($\alpha = 0,05$)
b. Estimar el parámetro de interés con una confianza del 95%.
c. ¿Podría decir cuál es la mejor ración? Justifique estadísticamente de acuerdo sus resultados anteriores.

Rtas: (a) Prueba de hipótesis de comparación de medias apareadas. Valor t = -2.74, valor p = 0.0289. Se rechaza la hipótesis nula. (b) -5.25 ± 4.53;(c) La ración B produce mejores resultados que la ración A. Esto queda justificado por el valor p de la prueba y por los extremos del IC.

[18] Se desea poner a prueba si el tipo de labranza influye sobre el nivel de malezas de los lotes. Para ello, se tomó una muestra aleatoria de 184 lotes y se los clasificó según el tipo de labranza (siembra directa, labranza convencional o labranza vertical) y el nivel de malezas (alto, medio, bajo); los resultados se observan en la siguiente tabla.

		Nivel de Malezas			
		Alto	Medio	Bajo	
Tipo de Labranza	Directa	28	22	16	66
	Vertical	22	22	18	62
	Convencional	12	20	24	56
		62	64	58	184

¿Existe relación entre el tipo de labranza y el nivel de malezas? Use $\alpha = 0.05$.

Prueba de Independencia; χ^2 calculado: 7.63; p = 0.1061; G de L = 4, α = 0.05, χ^2 tabla: 9.4877. Valor $p > \alpha$, entonces no se rechaza H_0. No hay evidencia de que el nivel de malezas sea dependiente del tipo de labranza.

[19] Al finalizar un curso de asistencia no obligatoria, un profesor realizó la siguiente agrupación basada en la aprobación o no del curso y la asistencia al mismo. Usando α = 0.05, ¿a qué conclusión puede llegar? ¿Cuál es el valor p?

Número de días ausente	Nota en el Curso	
	Aprobado	Reprobado
0-3	84	5
4-6	60	8
Más de 6	10	25

Prueba de Independencia. χ^2 calculado: 72.81; $p = 1.5465 \cdot 10^{-16}$; G de L = 2, α = 0.05, χ^2 tabla: 5.9914. Valor $p < \alpha$; entonces se rechaza H_0. Por lo tanto la calificación no es independiente de la asistencia.

[20] Una empresa de agroquímicos sabe por datos históricos que durante el verano la venta de sus productos se distribuye de la siguiente manera: 60 % herbicidas, 30 % fungicidas y 10 % de otros compuestos. Durante el verano del 2005 se registran las siguientes ventas: 100 corresponden a herbicidas, 15 a funguicidas, y 20 a otros productos. ¿Las ventas del verano de este año están en concordancia con los datos históricos? Utilice α = 0.05.

Prueba de bondad de ajuste. χ^2 calculado: 23.64; $p = 7.3489 \cdot 10^{-6}$; G de L = 2, α = 0.05, χ^2 tabla: $\chi^2_{2;0.05} = 5.9915$. Valor $p < \alpha$; entonces se rechaza H_0. Por lo tanto las ventas de verano del 2005 no coinciden con lo esperado según datos históricos.

[21] Una revista agropecuaria dispone de datos suministrados por varias empresas que fabrican y distribuyen agroquímicos sobre sus ventas y los gastos incurridos por cada empresa en publicidad en esa revista. Ambas variables están expresadas en pesos. La siguiente tabla resume los resultados obtenidos en un análisis de regresión lineal sobre estas variables:

Variable	N	R^2
Ventas	31	0.94

Matriz de coeficientes de regresión

Coef.	Est.	E.E.	LI(95%)	LS(95%)	T	p
Interc.	-19212.74	15251.77	-50406.10	11980.62	-1.26	0.22
Pendiente	1.76	0.08	1.58	1.93	20.78	0.00

a. Escriba la ecuación ajustada correspondiente y describa las estimaciones de los parámetros en términos del problema. ¿Datos sobre cuántas empresas fueron considerados en este análisis?

$-19212.74 + 1.76 \cdot x$; -19212.74 es la ordenada al origen, en otras palabras el volumen de ventas estimado cuando el gasto en publicidad es igual a 0. Aunque en este caso no tiene significado práctico. 1.76 es la pendiente o sea el cambio en ventas por cada unidad de gasto (pesos) en publicidad realizado.

b. Usted es gerente de una empresa de agroquímicos, considerando su respuesta en a) ¿decidiría invertir en publicar avisos en esta revista?. Justifique su respuesta.

La pendiente es significativa y positiva. Los gastos en publicidad explican el 94 % del volumen de ventas. Por lo tanto decidiría invertir en publicidad en esta revista.

c. Calcule el valor de ventas estimado para una empresa que invirtió 200 000 $ en publicidad en esta revista (asuma que este valor está dentro del rango de estimación posible del modelo).

$\hat{y}$ = 332787.26 $, para x = 200 000.

[22] La Secretaría de Agricultura y el Ministerio de Economía están interesados en determinar cuánto será el rendimiento de maíz en la localidad de San Lorenzo en el año 2004. Se dispone de los siguientes datos sobre rendimiento (en quintales por ha) y precipitaciones (en mm) desde 1992 hasta 2001.

Rendmiento	78	91	85	62	85	88	112	46	106	66
Precipitaciones	1328	1289	1371	1401	1350	1271	1215	1517	1285	1431

a. Ajuste un modelo lineal entre ambas variables. ¿Cuál es la variable dependiente y cuál la independiente?

modelo lineal $365.86 - 0.21 \cdot x$; Variable dependiente: *rendimiento*, Variable independiente: *precipitaciones*;

b. Describa los parámetros incluidos en el modelo e incluya las unidades en las que deben ser expresados. ¿Que significa en términos del problema que exista una pendiente negativa?

365.86 es la ordenada al origen, el rendimiento estimado cuando la precipitación es igual a 0, aunque no tiene significado biológico. 0.21 quintales$\cdot$ha^{-1}/ mm es la pendiente o sea el cambio en rendimiento por cada unidad (mm) de precipitación. Una pendiente negativa indica que a medida que la precipitación aumenta, el rendimiento disminuye.

c. San Lorenzo es una zona de elevada precipitación que es frecuentemente afectada por inundaciones. Se espera que el 2004 sea un año relativamente húmedo con 1500 mm de precipitaciones ¿cuál seria el rendimiento esperado?.

rendimiento para x = 1500 mm: 50.86 quintales ha^{-1}

[23] En la siguiente tabla se detalla la *inversión* hecha y la *ganancia* obtenida en miles de pesos para 12 explotaciones agropecuarias en la prov. de Buenos Aires durante el año 2002:

inversión	16	11	14	16	18	20	31	14	20	19	11	15
ganancia	5	2	3	5	3	7	10	6	10	5	6	6

a. Presente la estimación del modelo de regresión lineal para predecir la ganancia esperada en función de la inversión de capital hecha.

$0.17 + 0.32 \cdot x$;

b. ¿Presenta este modelo suficiente evidencia a un nivel de significación de 0.05 de que la ganancia en las explotaciones agropecuarias de la provincia de Bs. As. está determinada por la inversión realizada? Justifique su respuesta.

Si, presenta suficiente evidencia. $R^2 = 0.47$, $p = 0.01$;

c. Según el modelo propuesto ¿que ganancia se espera para una inversión de 30000 $?

Ganancia esperada para $x = 30$: 9.77 miles de $;

d. Usando el modelo propuesto sería correcto predecir la ganancia esperada para un establecimiento que invierte 100000$. Justifique su respuesta.

No es correcto predecir la ganancia esperada para $x = 100$ porque cae fuera del rango de predicción del modelo.

ESTADISTICA GENERAL: MODELO DE EXAMEN FINAL

- Este examen contiene 15 preguntas con 6 respuestas propuestas cada una. Identificar y marcar la única respuesta correcta en cada caso.
- Se aprueba con 9 repuestas correctamente identificadas.
- Tiempo disponible: 2 horas.

NORMAS

1. PARA RESOLVER ESTE EXAMEN, LOS ALUMNOS PUEDEN UTILIZAR UNA CALCULADORA CIENTÍFICA DE BOLSILLO, LAS TABLAS DE PROBABILIDADES PROVISTAS POR LA CATEDRA Y HASTA UNA HOJA CON FORMULAS COMO AYUDA MEMORIA.
2. **ESTE EXAMEN ES ESTRICTAMENTE INDIVIDUAL.**

CUALQUIER INFRACCION DE ESTAS NORMAS RESULTARA EN LA ANULACION INMEDIATA DEL EXAMEN PARA EL O LOS ALUMNOS INVOLUCRADOS.

NOMBRE Y APELLIDO: ______________________________

1. La emisión de gases contaminantes por los escapes de los autos es una de las principales formas de contaminación de la atmósfera. Los autos emiten tanto gases como el dióxido de carbono, que contribuyen al calentamiento global de la atmósfera, como substancias altamente patógenas como los hidrocarburos y el monóxido de carbono. Para poner a prueba un dispositivo de control de emisiones de monóxido de carbono, se tomaron al azar 10 autos de la ciudad de Buenos Aires y se midió su nivel de emisión antes y después de la instalación del dispositivo en cuestión. Los datos obtenidos son los siguientes.

	Emisión de monóxido de carbono (g/km)									
Auto	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
sin dispositivo de control	6,3	10,2	10,1	14,0	7,8	11,4	14,5	12,2	6,1	11,1
con dispositivo de control	5,5	8,1	9,2	7,0	8,5	6,8	10,3	5,4	2,5	8,2

1.1. ¿Cuál es un estimador insesgado de la varianza de la diferencia en la emisión de dióxido de carbono entre autos de la ciudad de Buenos Aires con y sin el dispositivo bajo prueba?

a. 2,60 g/km O
b. 6,45 g^2/km^2 ●
c. 6,79 g^2/km^2 O
d. 5,06 g/km O
e. 2,60 g^2/km^2 O
f. 5,06 g^2/km^2 O

1.2. ¿Cuál de los siguientes es el menor nivel de significación que conduce a aceptar la hipótesis que dice: *El dispositivo bajo prueba reduce el nivel esperado de emisión de monóxido de carbono de un taxi*?

a. α = 3,5839 O
b. α = 0,9877 O
c. α = 0,0250 O
d. α = 0,0100 O
e. α = 0,0050 ●
f. α = 0,0005 O

1.3. ¿Cuál error puede cometerse cuando se acepta la hipótesis referida en el punto anterior?

a. Tomar por ineficaz a un dispositivo efectivo O
b. Tomar por efectivo a un dispositivo sin efecto ●
c. Error de tipo II O
d. Tomar por ineficaz a un dispositivo sin efecto O
e. Tomar por efectivo a un dispositivo efectivo O
f. Dar la prueba por no concluyente O

2. La palmera *Butia yatay* produce frutos carnosos por fuera y leñosos por dentro denominados drupas. La mayoría de las drupas tienen una sola semilla (uniseminadas) pero un 20 % de ellas tienen dos semillas en el interior (biseminadas).

2.1. ¿Cuál es el valor esperado del número de semillas por drupa?

a. 0,4	O	d. 1,6	O
b. 0,8	O	e. 2,0	O
c. 1,2	●	f. 2,4	O

2.2. Si se toma una muestra aleatoria de 10 drupas ¿Cuál es la probabilidad de que a lo sumo dos de ellas sean biseminadas?

a. 0,302	O	d. 0,678	●
b. 0,208	O	e. 0,20	O
c. 0,040 y 0,168	O	f. 1	O

2.3. Si se toma una muestra integrada por 2 drupas elegidas al azar ¿Cuál es la probabilidad de que una sea uniseminada y la otra biseminada?

a. 0,16	O	d. 0,32	●
b. 0,20	O	e. 0,80	O
c. 0,20 y 0,80	O	f. 1	O

3. Una compañía productora de semilla afirma que, en la Pampa Ondulada, el rendimiento esperado de los cultivos de su híbrido de maíz genéticamente modificado es de 9,94 tn/ha con una varianza igual a 0,25 tn^2/ha^2. Suponiendo que lo que afirma la compañía fuera correcto y que el rendimiento de los cultivos de dicho híbrido fuera una variable aleatoria con distribución normal.

3.1. ¿Qué podría causar la varianza de los rendimientos?

a. Todos los productores aplican fertilizante.	O	d. Se trata de un cultivar genéticamente modificado.	O
b. Los suelos de la región han perdido fertilidad.	O	e. La sequía de verano limita el crecimiento de las plantas.	O
c. Algunos cultivos son sembrados más temprano y otros más tarde en el año.	O	f. El ambiente de la Pampa Ondulada es óptimo para la producción de maíz.	O

3.2. ¿Cuál sería la probabilidad de que el rendimiento de un cultivo de este híbrido tomado al azar superara los 10000 kg/ha?

a. 0,0500	O	d. 0,5478	O
b. 0,4052	O	e. 0,5948	O
c. 0,4522	●	f. 1,0088	O

3.3. ¿Cuál sería la probabilidad de que la media aritmética de los rendimientos de 4 cultivos tomados al azar fuera menor que 9500 kg/ha?

a. 0,0392	●	d. 0,8925	O
b. 0,1075	O	e. 0,9608	O
c. 0,6915	O	f. 1,546	O

4. En una pastura de 15 has se distribuyeron al azar 16 parcelas de 1 m². Todo el forraje presente dentro de cada parcela fue cortado, secado y pesado. Con los valores de los pesos obtenidos $y_i \ (i = 1,\ldots,16)$, se calcularon los siguientes estadísticos:

$$\bar{y} = \frac{1}{16}\sum_{i=1}^{16} y_i = 412\ g \qquad \text{y}$$

$$s^2 = \frac{1}{15}\sum_{i-1}^{16} (y_i - \bar{y})^2 = 9216\ g^2$$

4.1. ¿Qué es el valor $s^2 = 9216\ g^2$?

a. El valor esperado del cuadrado de la diferencia entre el peso del forraje de una parcela tomada al azar y la media poblacional. O

b. Un estimador insesgado de la varianza de los pesos de forraje entre todas las parcelas de 1m² de la pastura. ●

c. Un estimador insesgado de la varianza de los pesos de forraje en la muestra. O

d. El promedio de los cuadrados de los desvíos entre los pesos del forraje observados y la media muestral. O

e. La suma de los cuadrados de las diferencias entre los pesos del forraje observados y la media muestral. O

f. La varianza de los pesos de forraje entre todas las parcelas de 1m² de la pastura. O

4.2. Los límites de un intervalo del 95% de confianza para el peso <u>total</u> del forraje presente en la pastura son:

a. l = 36,1 tn; u = 46,3 tn O

b. l = 55,1 tn; u = 68,1 tn O

c. l = 360,8 g/m²; u = 463,2 g/m² O

d. l = 3,6 tn/ha; u = 4,6 tn/ha O

e. l = 369,9 g/m²; u = 454,1 g/m² O

f. l = 54,1 tn; u = 69,5 tn ●

4.3. ¿Qué significa el intervalo de confianza construido?

a. La probabilidad de que este intervalo contenga a la media muestral es de 0,95. O

b. Podemos tener un grado de confianza de 0,95 en que el intervalo contiene al peso total de forraje. ●

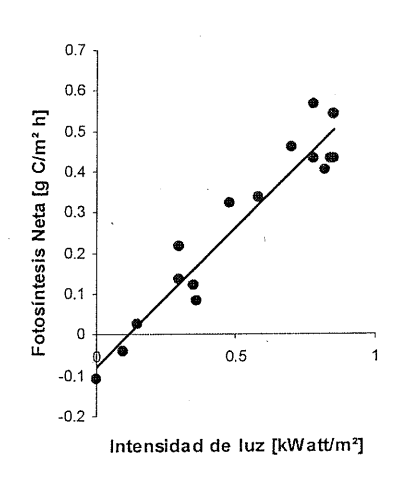
c. Es un intervalo que contiene al peso total de forraje. O

d. Podemos tener un grado de confianza de 0,95 en que la media muestral está dentro del intervalo. O

e. La probabilidad de que este intervalo contenga al peso total de forraje es 0,95. O

f. Es un intervalo que contiene a la media poblacional. O

5. En un estudio de fisiología vegetal se evaluó la fotosíntesis neta de cultivos de trigo (en gramos de Carbono fijado por m^2 y por hora) sometidos a diferentes niveles de intensidad de luz (en $kWatt/m^2$) asignados al azar. Cuando la fotosíntesis neta es positiva, el cultivo gana carbono (crece) y cuando es negativa pierde carbono (se reduce). A continuación se muestra un gráfico de dispersión y los principales resultados de un análisis de regresión lineal simple realizado con los datos obtenidos.



Análisis de regresión lineal

Variable	N	R^2
Int. de luz	16	0,9148

Coeficientes de regresión y estadísticos asociados

Coef	Est.	EE
Const	-0,0809	0,0330
Int. de luz	0,6863	0,0560

5.1. ¿Qué unidades tiene β_0?

a. $\frac{kWatt}{m^2}$ O

b. $\frac{g\,C}{m^2 h}$ ●

c. $\frac{kWatt\; g\,C}{m^2 h}$ O

d. $\frac{g\,C}{m^2 h} \Big/ \frac{kWatt}{m^2}$ O

e. $\frac{g\,C}{kWatt\; h}$ O

f. no tiene unidades O

5.2. ¿Cual de los siguientes es el menor nivel de significación con el cual es rechazada la hipótesis nula $\beta_0 \geq 0$?

a. 0,005 O

b. 0,01 O

c. 0,025 ●

d. 0,05 O

e. 0,95 O

f. 2,45 O

5.3. ¿Qué significa la hipótesis nula puesta a prueba en el punto anterior?

a. En la oscuridad, la fotosíntesis neta promedio de los cultivos de trigo es negativa. O

b. A medida que disminuye la intensidad de luz los cultivos de trigo pierden carbono. O

c. En promedio, los cultivos de trigo pierden carbono en la oscuridad O

d. Existe una relación estadística positiva entre la fotosíntesis neta y la intensidad de luz O

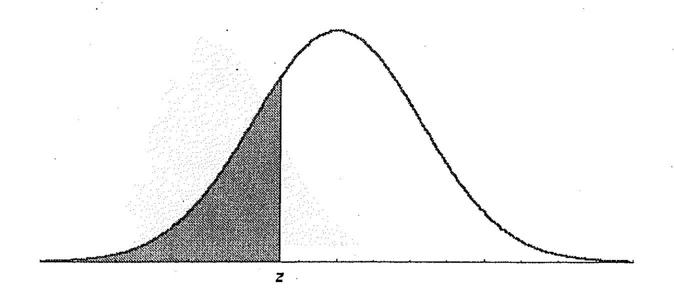
e. La fotosíntesis neta promedio aumenta con la intensidad de luz. O

f. En promedio, los cultivos de trigo no pierden carbono en la oscuridad ●

Respuestas Correctas:

1.1 b, 1.2 e, 1.3 b, 2.1 c, 2.2 d, 2.3 d, 3.1 c, 3.2 c, 3.3 a, 4.1 b, 4.2 f, 4.3 b, 5.1 b, 5.2 c, 5.3 f.

Probabilidades de cola izquierda de la distribución normal Standard. Valores de *z* negativos

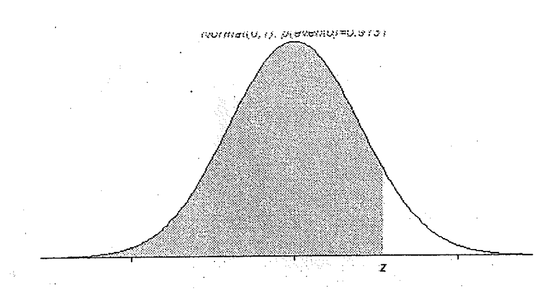


z	9	8	7	6	5	4	3	2	1	0
-3.0	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013
-2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019
-2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026
-2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035
-2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047
-2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062
-2.4	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082
-2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107
-2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139
-2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179
-2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228
-1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287
-1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359
-1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446
-1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548
-1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668
-1.4	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808
-1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968
-1.2	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151
-1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357
-1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587
-0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841
-0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119
-0.7	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420
-0.6	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743
-0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085
-0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446
-0.3	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821
-0.2	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207
-0.1	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602
0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000

Ejemplo de uso:

Para encontrar la probabilidad de que z sea menor a -1.36 encuentre en las filas el valor -1.3 y en las columnas el valor 6. La probabilidad de esa celda es el valor buscado. $P(z < -1.36) = 0.0869$.

Probabilidades de cola izquierda de la distribución normal Standard.
Valores de z positivos

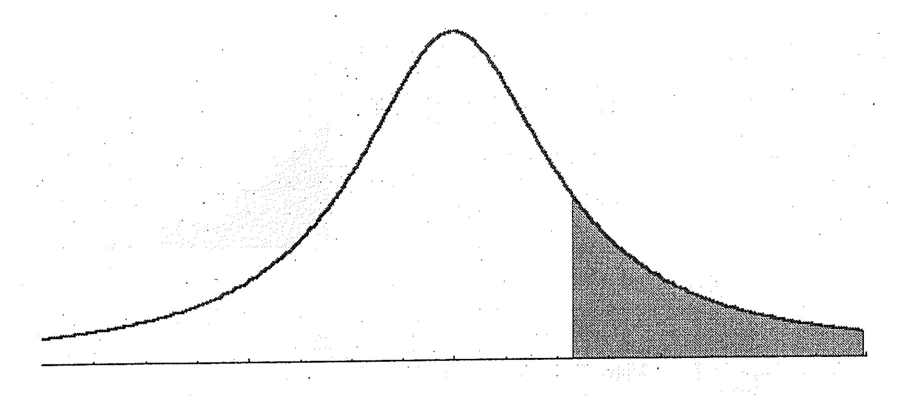


z	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Ejemplo de uso:

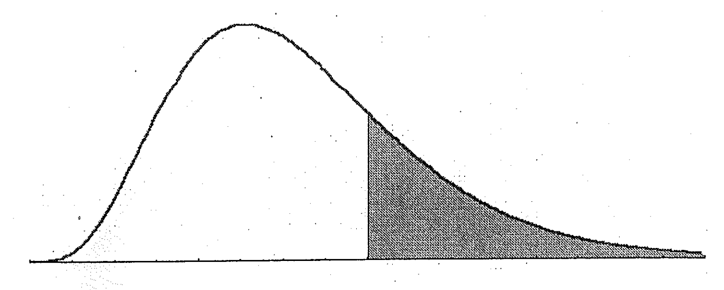
Para encontrar la probabilidad de que z sea menor a +1.36 encuentre en las filas el valor 1.3 y en las columnas el valor 6. La probabilidad de esa celda es el valor buscado. $P(z < +1.36) = 0.9131$. Si se desea encontrar la probabilidad de que z sea mayor a +1.36, reste a 1 la probabilidad encontrada anteriormente: $P(z > +1.36) = 1 - 0.9131 = 0.0869$.

Probabilidades de cola derecha de la distribución *t* de Student



	Probabilidad												
g.l.	**0.45**	**0.40**	**0.35**	**0.30**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.705	31.821	63.657	636.619
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.995	9.925	31.598
3	0.137	0.277	0.424	0.584	0.765	0.987	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.809
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.800	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.126	0.254	0.387	0.527	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

Probabilidades de cola derecha de la distribución χ^2



	Probabilidad						
g.l.	0.200	0.150	0.100	0.050	0.025	0.010	0.005
1	1.64	2.07	2.71	3.84	5.02	6.63	7.88
2	3.22	3.79	4.61	5.99	7.38	9.21	10.60
3	4.64	5.32	6.25	7.81	9.35	11.34	12.84
4	5.99	6.74	7.78	9.49	11.14	13.28	14.86
5	7.29	8.12	9.24	11.07	12.83	15.09	16.75
6	8.56	9.45	10.64	12.59	14.45	16.81	18.55
7	9.80	10.75	12.02	14.07	16.01	18.48	20.28
8	11.03	12.03	13.36	15.51	17.53	20.09	21.95
9	12.24	13.29	14.68	16.92	19.02	21.67	23.59
10	13.44	14.53	15.99	18.31	20.48	23.21	25.19
11	14.63	15.77	17.28	19.68	21.92	24.73	26.76
12	15.81	16.99	18.55	21.03	23.34	26.22	28.30
13	16.98	18.20	19.81	22.36	24.74	27.69	29.82
14	18.15	19.41	21.06	23.68	26.12	29.14	31.32
15	19.31	20.60	22.31	25.00	27.49	30.58	32.80
16	20.47	21.79	23.54	26.30	28.85	32.00	34.27
17	21.61	22.98	24.77	27.59	30.19	33.41	35.72
18	22.76	24.16	25.99	28.87	31.53	34.81	37.16
19	23.90	25.33	27.20	30.14	32.85	36.19	38.58
20	25.04	26.50	28.41	31.41	34.17	37.57	40.00
21	26.17	27.66	29.62	32.67	35.48	38.93	41.40
22	27.30	28.82	30.81	33.92	36.78	40.29	42.80
23	28.43	29.98	32.01	35.17	38.08	41.64	44.18
24	29.55	31.13	33.20	36.42	39.36	42.98	45.56
25	30.68	32.28	34.38	37.65	40.65	44.31	46.93
26	31.79	33.43	35.56	38.89	41.92	45.64	48.29
27	32.91	34.57	36.74	40.11	43.19	46.96	49.65
28	34.03	35.71	37.92	41.34	44.46	48.28	50.99
29	35.14	36.85	39.09	42.56	45.72	49.59	52.34
30	36.25	37.99	40.26	43.77	46.98	50.89	53.67
40	47.27	49.24	51.81	55.76	59.34	63.69	66.77
50	58.16	60.35	63.17	67.50	71.42	76.15	79.49
60	68.97	71.34	74.40	79.08	83.30	88.38	91.95
70	79.71	82.26	85.53	90.53	95.02	100.43	104.21
80	90.41	93.11	96.58	101.88	106.63	112.33	116.32
90	101.05	103.90	107.57	113.15	118.14	124.12	128.30
100	111.67	114.66	118.50	124.34	129.56	135.81	140.17